465

D-S
~~2123~~

NAVAL RESEARCH LOGISTICS QUARTERLY

MARCH 1970
VOL. 17, NO. 1

2 AUG 70

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

Rear Admiral H. E. Eccles, USN (Retired)
The George Washington University

O. Morgenstern
Princeton University

F. D. Rigby
Texas Technological College

D. M. Gilford
U.S. Office of Education

S. M. Selig
Managing Editor
Office of Naval Research
Washington, D.C. 20360

ASSOCIATE EDITORS

R. Bellman, RAND Corporation
J. C. Busby, Jr., Captain, SC, USN (Retired)
W. W. Cooper, Carnegie-Mellon University
J. G. Dean, Captain, SC, USN
G. Dyer, Vice Admiral, USN (Retired)
P. L. Folsom, Captain, USN (Retired)
M. A. Geisler, RAND Corporation
A. J. Hoffman, International Business Machines Corporation
H. P. Jones, Commander, SC, USN (Retired)
S. Karlin, Stanford University
H. W. Kuhn, Princeton University
J. Laderman, Office of Naval Research
R. J. Lundegard, Office of Naval Research
W. H. Marlow, The George Washington University
B. J. McDonald, Office of Naval Research
R. E. McShane, Vice Admiral, USN (Retired)
W. F. Millson, Captain, SC, USN
H. D. Moore, Captain, SC, USN (Retired)
M. I. Rosenberg, Captain, USN (Retired)
D. Rosenblatt, National Bureau of Standards
J. V. Rosapepe, Commander, SC, USN (Retired)
T. L. Saaty, University of Pennsylvania
E. K. Scofield, Captain, SC, USN (Retired)
M. W. Shelly, University of Kansas
J. R. Simpson, Office of Naval Research
J. S. Skoczylas, Colonel, USMC
S. R. Smith, Naval Research Laboratory
H. Solomon, The George Washington University
I. Stakgold, Northwestern University
E. D. Stanley, Jr., Rear Admiral, USN (Retired)
C. Stein, Jr., Captain, SC, USN (Retired)
R. M. Thrall, University of Michigan
C. B. Tompkins, University of California
J. F. Tynan, Commander, SC, USN (Retired)
T. C. Varley, Office of Naval Research
J. D. Wilkes, Department of Defense, OASD (IS/

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics ar will publish research and expository papers, including those in certain areas of mathematics, statistics, and economic relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, Jun September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printir Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individu issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this quarterly are those of the authors and not necessarily those of the Offi of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulatior NAVSO P-35.

Permission has been granted to use the copyrighted material appearing in this publication.

MACHINE SEQUENCING:
DISJUNCTIVE GRAPHS AND DEGREE-CONSTRAINED SUBGRAPHS*

Egon Balas

Carnegie-Mellon University

ABSTRACT

In two earlier papers, we proposed algorithms for finding an optimal sequence of processing m items on q machines, by finding a minimaximal path in a disjunctive network. In a third paper, this latter model was generalized (from 2-state to 3-state disjunctive graphs) so as to accommodate project scheduling with resource constraints. In this paper, we discuss another algorithm for the (2-state) disjunctive network problem, closely related to those mentioned above. To make the paper self-contained, section 2 briefly describes the problem. Section 3 introduces a class of constraints which forms the basis of the algorithm discussed in section 4. The constraints have only 1, −1, or 0 as coefficients on the left-hand side, integers on the right-hand side. The whole procedure of generating these constraints and finding a feasible solution whenever a new constraint is added, can be interpreted (section 5) as a process of generating a graph with degree-constraints on its nodes, and then finding a subgraph satisfying the degree-constraints. The nodes of the graph are generated by solving a critical-path-problem, the feasible subgraphs are found by implicit enumeration.

1. INTRODUCTION

Machine sequencing [9–13, 15, 16, 18, 19] is one of the most frequently occurring real-world problems which, while theoretically "solved", cannot be handled for a realistic problem size by any of the existing optimization techniques. This motivates the attempts to find more efficient ones.

Starting from a model introduced by Roy and Sussman [18], in two earlier papers we have proposed algorithms [3, 4] for finding an optimal sequence of processing m items on q machines, by finding a minimaximal path in a disjunctive network. In a third paper [6], this latter model was generalized (from 2-state to 3-state disjunctive graphs) so as to accommodate project scheduling with resource constraints.

In this paper, we discuss another algorithm for the (2-state) disjunctive network problem, closely related to [3, 4]. To make the paper self-contained, section 2 briefly describes the problem. Section 3 introduces a class of constraints which forms the basis of the algorithm outlined in section 4. Section 5 interprets the algorithm in terms of degree-constrained subgraphs.

The algorithm discussed in this paper is a close relative of [3] and [4]. While [3] applies Benders' partitioning procedure [7] to a structured mixed-integer program, the method proposed in [4] applies the techniques of implicit enumeration [1] directly to the disjunctive arcs, without recourse to an integer program. The approach taken in this paper is midway between the two. Like in the mixed-integer programming techniques of [14] or [5], which combine the ideas of Benders' partitioning method with those of implicit enumeration, we generate constraints on a vector with $0-1$ components which characterizes the state of the search for an optimal subset of disjunctive arcs. The constraints have only

*The research underlying this paper was supported by the New York Scientific Center of IBM.

1, −1, or 0 as coefficients on the left-hand side, integers on the right-hand side. The whole procedure of generating these constraints and finding a feasible solution whenever a new constraint is added, can be interpreted as a process of generating a graph with degree-constraints on its nodes, and then finding a subgraph satisfying the degree-constraints. The nodes of the graph are generated by solving a critical-path-problem, the feasible subgraphs are found by implicit enumeration.

A related, but different, technique is described by Raimond [17].

2. PROBLEM FORMULATION

The "simple" machine sequencing problem is that of finding an optimal sequence for processing m items (or lots of items) on q machines, when each item has to be processed on a specified sequence of machines, i.e.

a) a given operation (job) on a given item has to be performed on a specified machine;

b) the operations pertaining to a given item have to be performed in a prescribed sequence;

c) there is freedom of choice as to the sequence of operations to be performed on a given machine; and

d) given the completion time for each operation, one is looking for a sequence minimizing total completion time (i.e., the time needed for processing all items).

Let all operations be indexed from 1 to n_m, the operations pertinent to item i being assigned the index set $N_i=\{n_{i-1}+1, \ldots, n_i\}$ (with $n_0=0$), and let us introduce two dummy operations: start (index 0), and end (index $n=n_m+1$). Let t_j be the time of starting operation j, and $d_{ij}\geq 0$ the minimum time lapse between the starting of operations i and j ($d_{0j}=0$, $j=n_{i-1}+1$, $i=1, \ldots, m$). If we are sequencing operations on individual items, then d_{ij} can be thought of as the completion time of operation i, plus possibly the set-up time for operation j. If the sequencing pertains to operations on lots of items, then d_{ij} is the required lead time for operation i in relation to j, again possibly including set-up time for j. Finally, let the operations pertinent to machine k be assigned the index set N^k, where $k\epsilon Q=\{1, \ldots, q\}$, and let $M=\{1, \ldots, m\}$, $N=\{0, 1, \ldots, n\}$. Obviously,

$$N=\{0\}\cup\{\underset{i\epsilon M}{\cup} N_i\}\cup\{n\}=\{0\}\cup\{\underset{k\epsilon Q}{\cup} N^k\}\cup\{n\}.$$

With these definitions, the ("simple"*) machine sequencing problem can be formulated as the problem P of finding real numbers t_j, $j=1, \ldots, n$, and

$$\min t_n, \tag{1}$$

$$t_{j+1}-t_j\geq d_{j,j+1}, \qquad j\epsilon\underset{i\epsilon M}{\cup}\{N_i-\{n_i\}\}, \tag{2}$$

$$t_n-t_{n_i}\geq d_{n_i,n}, \qquad i\epsilon M, \tag{3}$$

$$t_{n_{i-1}+1}\geq 0, \qquad i\epsilon M, \text{ and} \tag{4}$$

$$(t_h-t_j\geq d_{jh})\vee(t_j-t_h\geq d_{hj}), \qquad h, j\epsilon N^k, h\neq j;\ k\epsilon Q. \tag{5}$$

The precedence constraints (2) define the required sequence of operations for each item i, while the constraints (3), of the same type, simply express the requirement that the last operation on each

*"Simple" as opposed to the more general case in which hypothesis a) above is replaced by a'): a given operation on a given item has to be performed on any of the machines belonging to a specified set.

item must be finished before the whole process can be regarded as ended. On the other hand, the disjunctive pairs of constraints (5) translate the condition that each machine can process only one item at any given moment (i.e., for every two operations to be performed on the same machine, one has to be finished before the other one is started). It is easy to see that P always has a solution.

It is known that P is equivalent to the problem of finding a minimaximal path in a disjunctive graph [3]. We shall briefly recall the pertinent definitions.

Two arcs of a graph form a *disjunctive pair* if any path in the graph is allowed to meet at most one member of the pair. A graph containing disjunctive arcs is called a *disjunctive graph*, and denoted by

$$\mathscr{D} = (N; Z, \mathbb{W}), \tag{6}$$

where N is the set of nodes, Z the set of conjunctive (i.e., nondisjunctive), and $\mathbb{W}$ the set of (pairwise) disjunctive arcs.

Associating a node j with each variable t_j of P, a conjunctive arc $(j, j+1)$, (n_i, n), and $(0, n_{i-1}+1)$ with each constraint (2), (3), and (4), respectively, and a pair of disjunctive arcs $[(j, h), (h, j)]$ with each disjunctive pair of constraints (5), one obtains the disjunctive graph $\mathscr{D}$ associated with P. Figure 1 represents the disjunctive graph of a problem with five items and four machines ($N^1=\{1, 3, 6, 9\}$, $N^2=\{2, 5, 7\}$, $N^3=\{8, 10, 12\}$, $N^4=\{4, 11, 13\}$). The numbers d_{ij} now represent the length of the arcs (i, j).

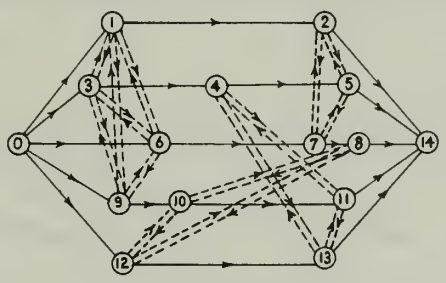


FIGURE 1

To be more specific, denote by $G=(N, Z)$ the graph associated with (2), (3), and (4). Then $\mathscr{D}$ is obtained from G by introducing the set $\mathbb{W}$ of pairwise disjunctive arcs, such that

$$\mathbb{W} = \bigcup_{k \in Q} \{(j, h) \in N^k \times N^k \mid j \neq h, \text{ and there is no path in } G \text{ connecting } j \text{ with } h\} \tag{7}$$

and

$$\left\{\begin{array}{l}(j, h) \text{ and } (r, s) \text{ form} \\ \text{a disjunctive pair}\end{array}\right\} \Longleftrightarrow \left\{\begin{array}{l}(j, h) \in \mathbb{W},\ (r, s) \in \mathbb{W} \\ \text{and } r=h,\ s=j\end{array}\right\}. \tag{8}$$

It should be mentioned here, that the above model can handle (see [4] for details) items resulting from the assembly of other items, as well as situations in which the operations pertaining to some item are only partially ordered.

A subset of $\mathbb{W}$ containing at most one arc of each disjunctive pair is called a *selection* (of disjunctive arcs). A selection $S_h \subset \mathbb{W}$ containing exactly one arc of each disjunctive pair is termed *complete*. Let

$$\mathscr{S} = \{S_1, \ldots, S_\sigma\} \tag{9}$$

be the set of complete selections. Obviously, if $p=\frac{1}{2}|\mathbb{W}|$ is the number of disjunctive pairs of arcs. then $\sigma=|\mathscr{S}|=2^p$.

Each complete selection $S_h \in \mathcal{S}$ generates a (conjunctive) graph of the form

$$G_h = (N, Z \cup S_h). \tag{10}$$

A *path* and a *circuit* in any of the graphs G_h is defined as usually (see, for instance, [8]), whereas the *length* of a path is the sum of the lengths of the arcs on the path. A *critical path* is a longest path from *source* (node 0) to *sink* (node n).

Let

$$\mathcal{G} = \{G_1, \ldots, G_\sigma\} \tag{11}$$

and

$$\mathcal{G}' = \{G_h \in \mathcal{G} \mid G_h \text{ has no circuits}\}. \tag{12}$$

A critical path in $G_k \in \mathcal{G}'$ is called a *minimaximal path in* $\mathcal{D}$ (and the associated selection S_k is called *optimal*), if, denoting by v_h the length of a critical path in G_h,

$$v_k = \min_{G_h \in \mathcal{G}'} v_h. \tag{13}$$

Problem P is then equivalent [3] to the problem of finding an optimal selection and an associated minimaximal path in the disjunctive graph $\mathcal{D}$.

3. A CLASS OF CONSTRAINTS

Let us partition the set W of disjunctive arcs into a subset W^+ of *normal* arcs and a subset W^- of *reverse* arcs, defined as

$$W^+ = \{(j, h) \in W \mid j < h\}, \qquad W^- = \{(j, h) \in W \mid h < j\}. \tag{14}$$

Obviously, $W = W^+ \cup W^-$, $W^+ \in \mathcal{S}$, $W^- \in \mathcal{S}$.

Each $S_h \in \mathcal{S}$ defines via (10) a graph $G_h \in \mathcal{G}$ which has a circuit or a critical path. A trivial, but rather lengthy way of finding a minimaximal path in $\mathcal{D}$, would be to enumerate all $S_h \in \mathcal{S}$, find a critical path (or a circuit) in each associated graph $G_h \in \mathcal{G}$, and retain the shortest critical path.

Now suppose we enumerate some but not all $S_h \in \mathcal{S}$, say those S_h such that $h \in H$, where H is a proper subset of the index set $\{1, \ldots, \sigma\}$ of $\mathcal{S}$. Also, suppose that in each associated graph G_h, $h \in H$, we find a critical path or a circuit C_h (we use the same notation for both cases), of length v_h. A disjunctive arc (i, j) of G_h will be called *critical* if and only if $(i, j) \in C_h$, where C_h is the critical path or circuit found in G_h (if there is more than one critical path or circuit, any one, but only one, can be used to define a set of critical disjunctive arcs). Let S_h^+ and S_h^-, respectively, denote the set of normal and reverse critical disjunctive arcs of G_h, i.e.,

$$S_h^+ = C_h \cap W^+, \qquad S_h^- = C_h \cap W^-. \tag{15}$$

Further, let $H' \subset H$ be the subset of indices $h \in H$ such that G_h has no circuits. Assume $H' \neq \emptyset$ (this is always the case if $W^+ \in H$ or $W^- \in H$). Let v^* denote the length of the shortest critical path in a graph belonging to the subset under consideration, i.e.,

$$v^* = \min_{h \in H'} v_h. \tag{16}$$

Finally, let us associate with each complete selection S_h a vector $y^h \in \{0, 1\}^p$, i.e., a vector with p zero-one components y_{ij}^h, one for each disjunctive pair of arcs $[(i, j), (j, i)]$, such that $y_{ij}^h = 0$ if S_h contains the normal arc, and $y_{ij}^h = 1$ if it contains the reverse arc of the pair:

$$y_{ij}^h = 0 \iff (i, j) \in S_h \cap W^+$$

(17)

$$y_{ij}^h = 1 \iff (j, i) \in S_h \cap W^-.$$

We shall call $y^h = (y_{ij}^h)$ the *characteristic vector* of S_h.

THEOREM 1: If there exists a complete selection $S_k \in \mathcal{S}$ such that the graph $G_k = (N, Z \cup S_k)$ has a critical path of length $v_k < v^*$, where v^* is defined by (16), then the characteristic vector y^k of S_k satisfies the constraints

(18) $$\sum_{(i,j) \in S_h^+} y_{ij}^k - \sum_{(j,i) \in S_h^-} y_{ij}^k \geq -|S_h^-| + 1, \quad \forall h \in H.$$

PROOF: Suppose (18) does not hold for some $r \in H$. Then $y_{ij}^k = 0$ for $(i, j) \in S_r^+$ and $y_{ij}^k = 1$ for $(j, i) \in S_r^-$, since this is the only set of values for which (18) is violated. But then all disjunctive arcs in C_r are also present in G_k. Hence, if C_r is a circuit, G_k also has a circuit; and if C_r is a critical path, then G_k either has a circuit, or has a critical path C_k of length $v_k \geq v_r \geq v^*$. In both cases the hypothesis of the theorem is contradicted.

Q.E.D.

4. THE ALGORITHM

The following two-phase algorithm for solving P uses Theorem 1. A sequence of complete selections $S_h \in \mathcal{S}$ is generated. Each S_h defines a graph $G_h = (N, Z \cup S_h)$. A critical path or a circuit C_h of G_h defines a constraint (18) on any further y to be considered. Since each y uniquely defines via (17) a complete selection, the inequalities (18) are in fact constraints on the selections to be generated. Suppose we start with the selection $S_1 = W^+$ and currently are at a stage when the selections $S_1, \ldots, S_{r-1}$ have already been generated. Let $\{1, \ldots, r-1\} = H$, and let v^* be defined by (16). A typical iteration of the algorithm then consists of the following two phases.

Phase I. Find y^r satisfying (18). If there is none, we are done: The selection S_k defined by the vector y^k associated with v^* is optimal, and a critical path C_k in $G_k = (N, Z \cup S_k)$ is a minimaximal path in $\mathcal{D}$. Otherwise, go to Phase II.

Phase II. Find a critical path or a circuit C_r in the graph $G_r = (N, Z \cup S_r)$, where S_r is defined via (17) by y^r. If C_r is a critical path such that $v_r < v^*$, update v^* (i.e., set $v^* = v_r$). In any case, generate a new constraint (18) and go to Phase I.

THEOREM 2. The algorithm consisting of Phases I and II ends in a finite number of iterations with (18) having no solution in $y \in \{0, 1\}^p$.

PROOF. $\mathcal{S}$ is finite, and no y^h (i.e., no $S_h \in \mathcal{S}$) can be repeated: each newly generated S_h is such that y^h satisfies (18); and each y^h is then used to generate a new constraint (18) which, by construction, is not satisfied by y^h.

Q.E.D.

There are several ways of implementing the above algorithm. We shall discuss a very simple implicit enumeration procedure for Phase I.

The repeated application of Phase I in successive iterations will be handled as a single process, in which the implicit enumeration is taken up at each iteration from where it had been left at the previous iteration.

The implicit enumeration proceeds through forward steps, exclusion tests, and backward steps (backtracking) [1, 2]. At any given stage of the process, each variable y_{ij} can be in one of the following three states: 1) y_{ij} has been assigned the value 1; 2) y_{ij} has been assigned the value 0; 3) y_{ij} has not been assigned any value. We shall refer to these three states by saying that y_{ij} is *fixed at 1*, *fixed at 0*, or *free*, respectively. A complete specification of the state of each component of the y vector is called a *pseudo-solution*.

The procedure starts with the pseudo-solution in which all y_{ij} are free, and generates a sequence of pseudo-solutions which can best be described in terms of an arborescence T. A node of T is associated with each pseudo-solution generated under the procedure (the root of T being the starting pseudo-solution). Each new node k is generated from some other node h by a *forward step*, which consists of fixing at 1 a variable y_{ij} which is free at node h, and leaving all other variables unchanged. The arc (h, k) of T is then associated with the variable y_{ij} (free at node h and fixed at 1 at node k). Since we are looking for a vector y satisfying (18), only those free variables are candidates for a forward step, which have a positive coefficient in at least one currently unsatisfied constraint. Various criteria can be used for selecting among the candidates the most appropriate free variable for a forward step. One possible criterion is the change in the length of a critical path we can obtain by setting to 1 a given free variable; a method for evaluating this change is given in [4].

Exclusion tests are meant to discover whether the set of descendants of a given node of T can possibly contain a node associated with a feasible y; in other words, whether by assigning arbitrary values 0 or 1 to the currently free components of y, one could possible satisfy all constraints (18). The easiest test of this type is to check whether the number of free variables with positive coefficient in a given constraint is less than the current (updated) value of the right-hand side (i.e., of the original right-hand side, plus the number of variables fixed at 1 and having negative coefficient, minus the number variables fixed at 1 and having positive coefficient in the given constraint). If so, then no set of 0 or 1 values assigned to the currently free variables can ever satisfy the constraint under examination, and hence no descendant of the current node needs to be examined; the node is abandoned and we step back (backtrack) to its immediate prececessor in T. On the other hand, in the particular case when a vector y obtained from the current pseudo-solution by setting all free components equal to 0 happens to satisfy all current constraints (18), we have a feasible y and go to Phase II.

When *backtracking*, we want to make sure that neither the abandoned node, nor any of its descendants will ever be visited in any future step of the procedure. This is achieved by fixing at 0 the variable associated with the backtracking arc.

The algorithm terminates when backtracking is due but impossible (since we are instructed to backtrack from the origin), i.e., when it is established that there is no $y \in \{0, 1\}^p$ satisfying the current set of constraints (18).

In this procedure, the constraints generated in Phase II are used to rule out (eliminate from the search) certain subsets of the y-set. While each Phase II ends by generating a new constraint which rules out the last y obtained in Phase I, the "old" (previously generated) constraints are used in Phase I to "screen" the set of y vectors from which the one for the next Phase II is selected. Omission of some of the "old" constraints does not affect the convergence of the procedure, since the partial enumeration of y vectors is in itself nonredundant. Hence we favor keeping only a fixed number of constraints. After

that number—say K—is attained, one of the "old" constraints should be dropped whenever a new one is generated. The constraint to be dropped should be chosen among those associated with one of the abandoned nodes of T; it might be the one which is currently most "oversatisfied."

The constraints (18) used in the above algorithm are not the strongest ones that one can generate by solving the current critical path problem. Their potential advantage lies in their simplicity: all coefficients are $+1$, -1, 0 on the left-hand side, integers on the right-hand side. It is this feature which makes possible an equivalent formulation in terms of a degree-constrained subgraph-problem, to be discussed in the next section.

5. DISJUNCTIVE GRAPHS AND DEGREE-CONSTRAINED SUBGRAPHS

We shall now associate with the disjunctive graph $\mathscr{D}$ a graph $\mathscr{B}=(\mathscr{X}, \mathscr{A})$ defined as follows.

The nodes of $\mathscr{B}$ belong to two disjoint subsets, i.e., $\mathscr{X}=\mathscr{X}_1 \cup \mathscr{X}_2$, $\mathscr{X}_1 \cap \mathscr{X}_2=\phi$. The set $\mathscr{X}_1$ contains p nodes, each one associated with a disjunctive pair of arcs in $\mathscr{D}$. The set $\mathscr{X}_2$ contains 2^p nodes, each one associated with a complete selection $S_h \in \mathscr{S}$. (The full set $\mathscr{X}_2$, however, need not be known: nodes in $\mathscr{X}_2$ are generated as the associated S_h are).

The arcs of $\mathscr{B}$ are defined as follows. Let j and h be distinct nodes of $\mathscr{B}$, such that $j \in \mathscr{X}_1$ and $h \in \mathscr{X}_2$. Let $[(j_1 j_2), (j_2, j_1)]$ be the disjunctive pair of arcs in $\mathscr{D}$ associated with node $j \in \mathscr{X}_1$ of $\mathscr{B}$, where $(j_1, j_2) \in W^+$; and let S_h be the selection of $\mathscr{D}$ associated with node $h \in \mathscr{X}_2$ of $\mathscr{B}$. Then $(j, h) \in \mathscr{A}$ if, and only if, $(j_1, j_2) \in S_h^+$; and $(h, j) \in \mathscr{A}$ if, and only if, $(j_2, j_1) \in S_{\bar{h}}$, where S_h^+ and $S_{\bar{h}}$ are defined by (15).

Thus, all arcs of $\mathscr{B}$ have one of their ends in $\mathscr{X}_1$ and the other one in $\mathscr{X}_2$ ($\mathscr{B}$ is bipartite).

The number of arcs having j as terminal node is called the *inward demi-degree* of j, and the number of arcs having j as initial node is called the *outward demi-degree* of j [8]. The difference of the two is the *degree* of j, denoted $d(j)$. By construction, the inward demi-degree of any $h \in \mathscr{X}_2$ is $|S_h^+|$, while its outward demi-degree is $|S_{\bar{h}}|$. Hence, for all $h \in \mathscr{X}_2$,

$$d(h)=|S_h^+|-|S_{\bar{h}}|. \tag{19}$$

With each $h \in \mathscr{X}_2$, we associate a real number

$$f(h)=-|S_{\bar{h}}|+1 \tag{20}$$

called the *degree-requirement* on h.

A subgraph G' of a graph G is defined [8] as a graph obtained from G by dropping some of the nodes of G and the arcs incident at those nodes. Hence G' contains all those arcs of G connecting nodes which are also nodes of G' (other definitions exist, but here we shall use this one). Thus G' is uniquely defined by the set of its nodes.

Let $\mathscr{B}'=(\mathscr{X}', \mathscr{A}')$ be a subgraph of $\mathscr{B}$, such that $\mathscr{X}'=\mathscr{X}_1 \cup \mathscr{X}_2'$, where $\mathscr{X}_2'$ is some subset of $\mathscr{X}_2$, and consider the following *degree-constrained subgraph-problem* (DCS-problem) associated with $\mathscr{B}'$:

Find a subgraph $\mathscr{B}''=(\mathscr{X}'', \mathscr{A}'')$ of $\mathscr{B}'$, such that

$$\mathscr{X}''=\mathscr{X}_1' \cup \mathscr{X}_2', \qquad \mathscr{X}_1' \subset \mathscr{X}_1, \tag{21}$$

and

$$d(h) \geq f(h), \ \forall h \in \mathscr{X}_2'. \tag{22}$$

The following relationship exists between our initial problem P and a given subgraph $\mathscr{B}'$ of $\mathscr{B}$ with its associated *DCS*-problem:

THEOREM 3: If $\mathscr{X}_2'$ is such that no subgraph $\mathscr{B}''$ of $\mathscr{B}'$ satisfying (21), (22) exists, then $\mathscr{X}_2'$ contains a node k associated with an optimal selection $S_k \in \mathscr{S}$.

PROOF: One can easily see that the *DCS*-problem associated with a given $\mathscr{B}'$ is equivalent to the Phase I problem of the algorithm described above, i.e., equivalent to finding a vector $y \in \{0,1\}^p$ satisfying the constraints (18), where the coefficient matrix of (18) is the node-to-node incidence matrix of the bipartite graph $\mathscr{B}'$, and the current index set H of (18) is the same as the index set $\mathscr{X}_2'$ of (22). In other words, there is a one-to-one correspondence between the nodes $k \in \mathscr{X}_1$ of $\mathscr{B}'$ and the variables y_{ij} of (18), as well as between the nodes $h \in \mathscr{X}_2'$ of $\mathscr{B}'$ and the constraints $h \in H$ of the set (18). Q.E.D.

Thus the algorithm described above can also be regarded as a method of generating a graph $\mathscr{B}'$ of the type defined above, by starting with $\mathscr{X}_2' = \emptyset$ and adding at each iteration one new node to $\mathscr{X}_2'$, until $\mathscr{X}_2'$ is such that $\mathscr{B}'$ has no subgraph $\mathscr{B}''$ satisfying (21), (22). At that point the current "best" selection $S_k \in \mathscr{S}$ is optimal.

A typical iteration of the algorithm can then be described in the following terms.

Phase I. Find a subgraph $\mathscr{B}''$ of the current graph $\mathscr{B}'$, satisfying the degree-constraints (22). If there is none, the procedure ends: the selection S_k, associated with the current v^*, is optimal. Otherwise, go the Phase II.

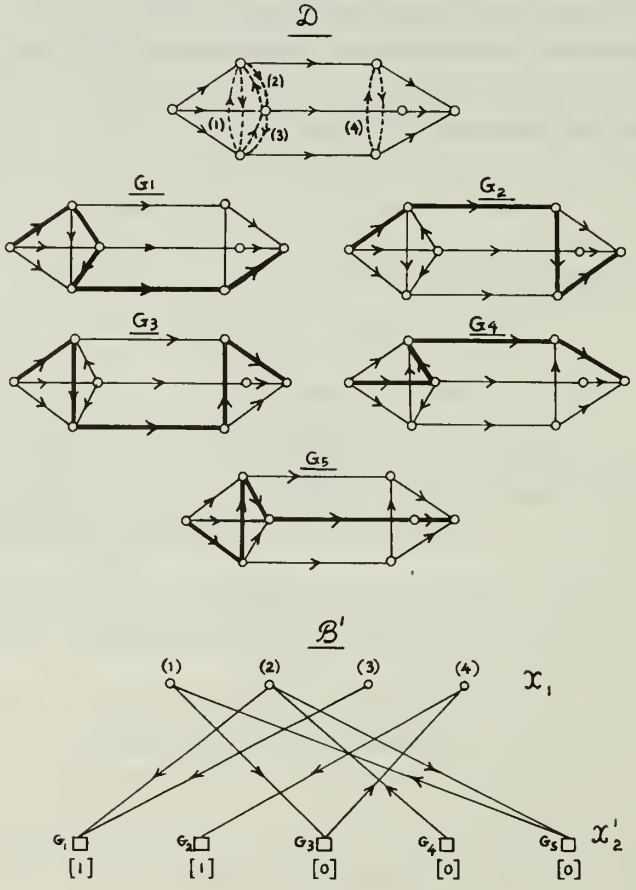


FIGURE 2

Phase II. Find a critical path or a circuit C_r in the graph $G_r = (N, Z \cup S_r)$, where S_r is defined by $\mathscr{B}'' = (\mathscr{X}'', \mathscr{A}'')$ according to the following rule: $(i, j) \epsilon S_r$ if the node of $\mathscr{B}'$ associated with the disjunctive pair $[(i, j), (j, i)]$ belongs to $\mathscr{X}_1'$; otherwise $(j, i) \epsilon S_r$. If C_r is a critical path such that $v_r < v^*$, update v^* (i.e., set $v^* = v_r$). In any case, generate a new node of $\mathscr{B}'$ associated with G_r, and the arcs of $\mathscr{B}'$ incident at the new node (these arcs are defined by C_r as shown at the beginning of this section), as well as the degree-requirement $f(r)$ defined by (20). Then go to Phase I.

Figure 2 illustrates this interpretation on a disjunctive graph $\mathscr{D}$ with four disjunctive pairs of arcs [indexed (1), (2), (3), and (4)]. A sequence of five graphs G_h, $h = 1, \ldots, 5$ is generated along with the associated graph $\mathscr{B}'$. The numbers in square brackets are the degree-requirements $f(h)$ associated with the nodes $h \epsilon \mathscr{X}_2'$. By examining $\mathscr{B}'$, one can see that the procedure is ended, since $\mathscr{B}''$ satisfying (21), (22) does not exist.

The discovery of an efficient method for solving the *DCS*-problem of Phase I would be of crucial importance to the whole area of machine-sequencing and related problems.

REFERENCES

[1] Balas, E., "An Additive Algorithm for Solving Linear Programs with Zero-One Variables," Operations Research *13*, 517–546 (1965).

[2] Balas, E., "Discrete Programming by the Filter Method," Operations Research *15*, 915–957 (1967).

[3] Balas, E., "Finding a Minimaximal Path in a Disjunctive PERT Network." Théorie des Graphes. Journées internationales d'étude, Rome, Juillet 1966 (Dunod (Paris)–Gordon and Beach, New York, 1967), pp. 21–30.

[4] Balas, E., "Machine Sequencing via Disjunctive Graphs: An Implicit Enumeration Algorithm," Operations Research *17*, No. 6 (1969).

[5] Balas, E., "Minimax and Duality for Linear and Nonlinear Mixed-Integer Programming," Integer and Nonlinear Programming (J. Abadie (Ed.), North-Holland, 1970).

[6] Balas, E., "Project Scheduling with Resource Constraints," Proceedings of the NATO Conference on the Applications of Mathematical Programming – June 24–28, 1968 (Cambridge).

[7] Benders, J. R., "Partitioning Procedures for Solving Mixed-Variables Programming Problems," Numerische Mathematik *4*, 238–252 (1962).

[8] Berge, C., Theory of Graphs (Methuen, London, 1966).

[9] Bowman, F. H., "The Schedule-Sequencing Problem," Operations Research *7*, 621–624 (1959).

[10] Conway, R. W., W. L. Maxwell, and L. W. Miller, Theory of Scheduling (Addison-Wesley, 1967).

[11] Dantzig, G. B., "A Machine-Job Scheduling Model," Management Science *6*, 191–196 (1960).

[12] Giffler, B. and G. L. Thompson, "Algorithms for Solving Production Scheduling Problems," Operations Research *8*, 487–503 (1960).

[13] Johnson, S. M., "Optimal Two and Three-Stage Production Schedules with Set-Up Times Included," Naval Research Logistics Quarterly *1*, 61–68 (1954).

[14] Lemke, C. L., and K. Spielberg, "Direct Search Algorithms for Zero-One and Mixed-Integer Programming," Operations Research *15*, 892–914 (1967).

[15] Manne, A. S., "On the Job-Shop Scheduling Problem," Operations Research *8*, 219–223 (1960).

[16] Németi, L., "Das Reihenfolgeproblem in der Fertigungsprogrammierung und Linearplanung mit logischen Bedigungen," Mathematika *6*(29), 87–99 (1964).

[17] Raimond, J. F., "An Algorithm for the Exact Solution of the Machine Scheduling Problem," IBM, New York Scientific Center Report No. 320–2930 (Feb. 1968).

[18] Roy, B., and B. Sussman, Les problèmes d'ordonnancement avec contraintes disjonctives. SEMA, Note D.S. No. 9 bis (décembre 1964).

[19] Wagner, H., "An Integer Programming Model for Machine Scheduling," Naval Research Logistics Quarterly *6*, 131–140 (1959).

* * *

REVIEW OF SEQUENCING RESEARCH

James E. Day

Chemical & Metallurgical Division
Sylvania Electric Products, Inc.
Towanda, Pennsylvania

and

Michael P. Hottenstein

College of Business Administration
The Pennsylvania State University
University Park, Pennsylvania

CLASSIFICATION OF SEQUENCING LITERATURE

The majority of research articles published on job shops appears to be concerned with the effects of scheduling and sequencing (dispatching) on various measures of shop performance criteria. Usually these effects are studied in a given context, a context which can be described by making the appropriate choice from each of the following three classifications:

1. Number of component parts comprising a job.
   a. Single-component jobs.
   b. Multi-component jobs which require assembly and/or subassembly operations.
2. Production factors possessed by the shop.
   a. Machines.
   b. Labor and machines.
3. Jobs available for processing.

a. N jobs to be scheduled, or sequenced, where N is finite. This is often referred to as the static scheduling or sequencing problem [65].

b. An undetermined (literally infinite) number of jobs arrive continuously, but randomly, at the shop for service. This is often referred to as the dynamic sequencing problem [65].

Thus (1a-2a-3a) would indicate the study of a shop situation in which N single-component jobs are to be sequenced through a shop with little or no regard for the effect of labor flexibility, absenteeism, etc. In the literature, only a few articles are concerned with multicomponent jobs and the labor factor. Most of the articles appear to be of the (1a-2a-3a) and (1a-2a-3b) variety.

Discussion of man-labor considerations and multicomponent jobs will be postponed until later. Attention will now be given to the (1a-2a-3a) and (1a-2a-3b) situations.

A schema for classifying sequencing problems is shown in Figure 1. The framework in Figure 1 shows that sequencing problems have been classified by: (1) the nature of job arrivals (i.e., fixed batch size vs continuous arrivals given by a probability density function); (2) the number of machines involved (i.e., single stage production, $M=1$, vs multistage production $M>1$); and (3) the nature of the job route. Further classification could be superimposed on Figure 1 such as setup times which are either dependent on or independent of job sequence on a given machine, and due date considerations. Regardless of the degree to which Figure 1 might be refined, it does serve to illustrate some of the many

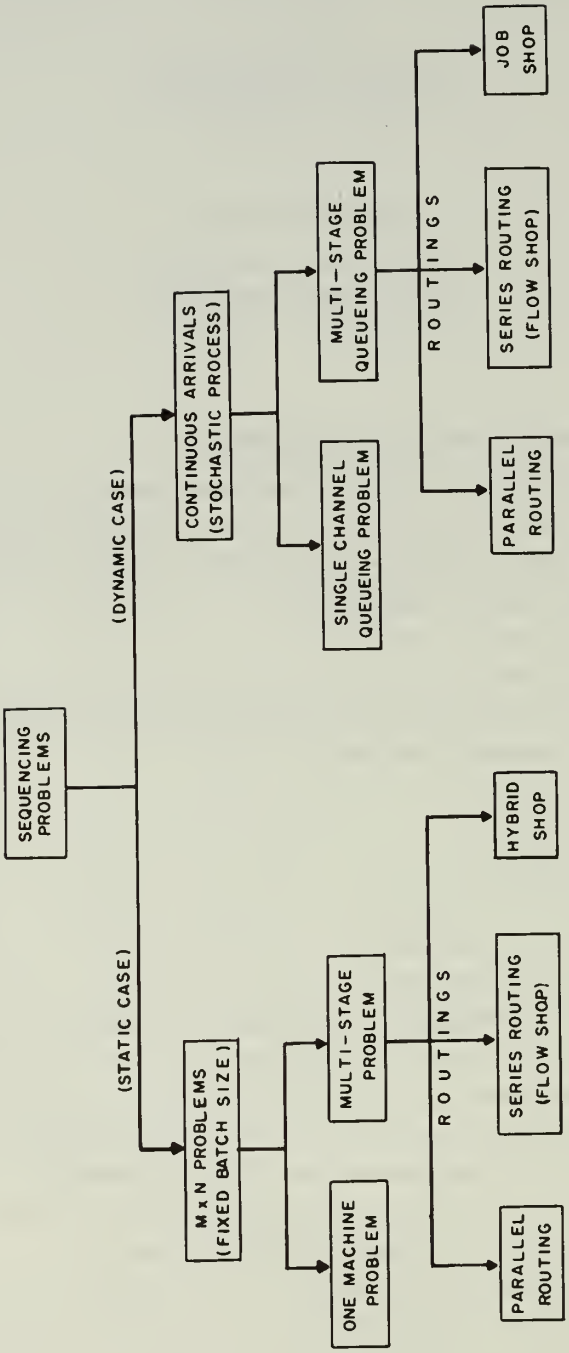


FIGURE 1. A classification of sequencing problems

facets of sequencing problems which have been considered in the literature. Note that both single and multicomponent jobs could be included under this framework; however, there is no consideration given to the effect of labor on sequencing.

STATIC SEQUENCING

The left main branch of Figure 1 represents a class of problems referred to as static $M \times N$ scheduling problems (scheduling and sequencing may be used interchangeably for this situation) [64, pg. 542]. All N jobs have simultaneous starting times in a shop composed of M machines ($M \geqq 1$). Usually there is a shop authorization function associated with $M \times N$ sequencing situations. Since orders arrive randomly at the shop, they may be held for some time, typically to level the shop load over a planning period, before being authorized for production. At that time a batch of orders may have accumulated whereupon they are released simultaneously to the shop for processing.

The Single Machine Case

In glancing at Figure 1, one immediately notes a separation of operational systems into single machines or processors (i.e., $M = 1$) and multiple processors ($M > 1$). The significance of studying the single machine case lies in the fact that it may shed light on the more complicated multiple machine problem. For example, the single machine problem has helped to render possible the investigation of optimization under several criteria [49]. Examples of single machine production systems include a food processing line, a high-speed computer, and an electric utility maintenance truck.

Research in the one machine static sequencing problem has been extensive. According to Elmaghraby [49], the optimal sequence has been found that minimizes the (1) maximum tardiness, (2) weighted sum of completion times, (3) weighted sum of tardiness, (4) total cost of tardiness, (5) total penalty if the jobs are "related" to each other, (6) total setup time or setup cost (when either is sequence-dependent), (7) total cost of processing when the processor is characterized by a single state-variable, (8) number of changeovers when the products are subject to a demand schedule, (9) total cost of production and inventory for a finite horizon, and (10) cost of changeovers when the cost is variable from product to product, but independent of sequence. Appropriate references are cited in his article.

For the one machine static sequencing rule, the shortest operation time priority rule (SOT), which is synonymous with the shortest processing time rule if setup times are not considered, has been found to be optimal for both due-date and non due-date criteria. The SOT rule, having been successfully applied in the single machine case, has also been shown to be an effective rule for the dynamic multiple machine case as will be discussed later.

The Study of Multiple Machine Systems

Referring again to Figure 1, consider now the static sequencing of multiple processors (i.e., $M > 1$). For this situation there are three possible arrangements of the machines in the system, namely, series, parallel, and hybrid. Static sequencing on each of these arrangements can be expressed by Gantt charts.

Static sequencing of N jobs on M machines in series (i.e., linear flow shop) has usually been optimized relative to the "makespan" criterion, the notable exception being an attempt towards optimization relative to the "average completion time" criterion for sequencing on two machines [88]. Literature concerning static sequencing on machines in series can be classified as follows:

1. Two jobs on M machines [2, 23, 74, 110, 157].*
2. N jobs on two machines [97, 98, 117, 118, 124, 158].

*The references are meant to be examples and are not necessarily exhaustive.

3. N jobs on three machines [17, 88, 97, 108, 156, 161].

4. N jobs on M machines [36, 37, 78, 122, 123, 125, 126, 152, 161].

For the parallel case of static sequencing there are N jobs available for processing at arbitrary time zero which must be processed exactly once on one of M machines [141]. Research to date has determined the optimum sequence which minimizes: (1) the cost of tardiness when the machines are identical, and (2) the total cost of production, which includes processing cost, setup cost, as well as the cost of tardiness on either similar or dissimilar machines, provided that the order in which the jobs are considered for analysis is known. If one wants to find the maximum number of machines to be purchased to perform a set of tasks, procedures are available for determining the optimal number required when the tasks are (1) immediately available, (2) available only at specified times and then continuously available thereafter, and (3) related to each other by precedence relationships [49]. All these results can be found in [40, 45, 46, 59, 104, 106, 115, 141, 142, 143].

Methods of Studying Static Sequencing

Four approaches in dealing with static sequencing are revealed by a survey of the literature, namely (1) combinatorial approaches, (2) general mathematical programming, (3) reliable heuristics, and (4) Monte Carlo sampling. Combinatorial approaches are based on the changing of one permutation to another by "switching around" of jobs that satisfy a given criterion. Literature references for combinatorial approaches include papers by Gapp, Mannkekar, and Mitten [62], Gilmore and Gomory [70], Jackson [89], Johnson [97, 98], McNaughton [115], Mitten [117], Root [141], and Smith [153].

General mathematical programming approaches to static sequencing include linear, dynamic, convex, and quadratic programming, integer programming, networks of flow, Lagrangian methods, and the like [49]. Though there are a great number of articles published on this subject, the following are representative: Balinski and Quandt [7], Bellman [12, 13], Bowman [15], Dantzig [34], Dantzig, Fulkerson, and Johnson [33], Elmaghraby [46, 48], Elmaghraby and Ginsberg [43], Glassey [72], Held and Karp [77], Manne [111], Rothkopf [142], Szwarc [157], Story and Wagner [156], and Wagner [161].

The third approach to static sequencing is one which employs reliable heuristics, also known as "combinatorial programming" or "controlled enumeration." This approach, according to Elmaghraby [49], is developed on the basis of two principal concepts: (1) the use of a controlled enumeration technique for considering all potential solutions, and (2) the elimination from explicit consideration of particular potential solutions which are known from dominance, bounding, and feasibility considerations to be unacceptable. Elmaghraby restricts the meaning of "combinatorial programming" to problem-solving procedures based on these concepts which are reliable in the sense that when carried to completion, they guarantee the discovery of an acceptable solution if one exists, or the knowledge that none exists.

An alternate name for such programs is "branch and bound." The idea of a "branch" stems from the fact that in terms of a tree of the problem the procedure attempts to select a branch of the tree to elaborate and evaluate. "Bound" emphasizes the effective use of means for bounding the value of the objective function at each node of the tree both for eliminating dominated parts and for choosing a branch for elaboration and evaluation.

References for the use of reliable heuristics as applied to static sequencing include Brown and Lomnicki [17], Eastman, Ever, and Issacs [40], Elmaghraby [47], Ignall and Schrage [88], Little, Murty, Sweeny, and Karel [107], Lomnicki [108], and McMahon and Burton [114].

Monte Carlo techniques have also been applied to static sequencing systems. For example,

consider the N job, M machine static sequencing problem. The maximum number of schedules which could be generated, if one allows the possibility of a different sequence on each machine, is $(N!)^M$. For large values of M and N, complete enumeration of all such schedules to find an optimum is impractical. By applying Monte Carlo sampling to the flow shop situation, Heller [78] has found that if processing time on each machine is assigned randomly from a probability density function uniformly distributed between zero and nine units of time, the resulting distribution of schedule times for a large number of experiments is normally distributed. Elmaghraby has used this result in considering the value and cost of continued sampling in the effort to obtain a shorter schedule time from the samples of sequences generated [49]. Further references on the use of Monte Carlo techniques in the solution of static sequencing problems can be found in Giffler and Thompson [66] and Giffler, Thompson, and V. Van Ness [68].

DYNAMIC SEQUENCING

Consider now the main right hand branch of Figure 1. Sequencing problems discussed in this branch will be referred to as dynamic due to the fact that jobs are constantly entering and leaving the job file according to some stochastic process.

The Single Channel Queueing Problem

Work on single channel (one machine) queueing problems has been quite extensive. Typical assumptions for this problem include: (1) Poisson arrival rates, (2) exponential service times, (3) first-come, first-served (FCFS) queue discipline, and (4) a mean service rate greater than the mean arrival rate of orders. The criteria by which the performance of this type of system might be evaluated are: (1) the expected length of the waiting line or the average number of jobs in the system, (2) the average job completion time, (3) machine utilization, and (4) the average waiting time for a job, which is the average completion time minus the average processing time [64, pg. 555]. For excellent treatment of the work done on single channel queues, the reader should refer to [32, chaps. 8 and 9; and 147].

Multi-Stage Queueing Problems

Finally, there are the dynamic multi-stage queueing problems which can be classified into three principle arrangements of machines [32, pg. 208]. Parallel-channel queues have a number of identical machines working in parallel to provide a single type of service. This corresponds to a machine center of several interchangeable machines. Tandem queues, or queues in series, correspond to the flow shop situation. Finally there is the general-network case (job shop) in which each order is processed by a particular subset of the machines in the shop.

With the assumptions of Poisson arrival rates, exponential service times, and the first-come, first-served queue discipline, the following quantities can be computed for parallel-channel queues: (1) mean number of jobs in a waiting line, (2) mean number of jobs in the system, (3) mean waiting time of a job, and (4) mean time (including service time) of a job in the system [19, pg. 257–258]. Other references on parallel-channel queues include [101] and [162].

In the flow shop situation machines (or machine groups) may be numbered so that every job routing specifies an increasing sequence of machine number identifiers. The input to a machine (machine group) consists of inputs from a dynamic job file or from machines (machine groups) with lower indices. It is also possible for a job to exist from the shop after being processed on any machine (machine group).

In comparison to the literature available on the dynamic job shop situation, relatively little has been published on dynamic flow shops. The interested reader is referred to [5], [27], [43], [61], [86], and [137].

THEORY AND RESEARCH OF DYNAMIC JOB SHOPS

More than 10 years ago, a large number of simulation experiments suggested the advantages to be gained by considering the job shop as a network of waiting lines with fixed short-run capacity, the short run being interpreted as one 8-hour shift. This recognition that the job shop could be represented as a system or network of queues was an important one, and it inspired a great deal of fundamental research which is still continuing. Evans [52, 53, 54, 55], Jackson [90, 91, 92, 95, 96], Nelson [128, 129, 130, 131], and others have treated the job shop as a network of queues. The early research on queueing suggested that an important variable which could be manipulated in sequencing models was the priority dispatching rule (queue discipline), and much of the research has focused upon comparative studies of dispatching rules through the use of computer simulation.

In the following discussion attention will be focused on literature of the (1a-2a-3b) variety. Neither labor-machine interaction nor multicomponent jobs will be considered.

Jackson's Decomposition Principle

To date the most powerful analytical result available in the study of dynamic job shops (and dynamic flow shops) is Jackson's decomposition principle [90]. In order to state the principle the following assumptions are made:

1. The arrival times for each job arriving from outside the system are exponentially distributed.
2. The processing times at each machine are exponentially distributed.
3. The jobs are routed to a machine by a fixed probability transition matrix. This matrix defines the probabilities of a job going from either machine l to any other machine in the shop or from machine l to the customer.
4. The priority rule at each machine is FCFS.

Given these assumptions, the system can be decomposed into one of a network of independent individual-machine queueing systems. This result was based on the fact that under FCFS, the output of a single queue is exponential if the interarrival and service times are exponential. In the dynamic job shop situation, the output of one machine becomes the input to the next in a network of queues.

Jackson's decomposition principle has been generalized to apply to systems on which the rate at which jobs are processed at each machine is an essentially arbitrary function of the queue length at that machine [94]. The lengths of the individual waiting lines again turn out to be independent, with distributions identical to those for similar isolated queueing systems.

Description of the Models Used in Studying Dynamic Job Shops

As with Jackson's decomposition principle, other studies of the dynamic job shop situation begin with certain assumptions. Typical assumptions include the following:

1. Arrival times and processing times are generated from stated probability density functions.
2. The mean arrival rate of jobs is less than the mean service rate of the shop.
3. Each operation, once started, must be performed to completion (i.e., no preemptive priorities).
4. Due dates are fixed.
5. Setup time for a given operation is included in the processing time of the job being serviced.
6. Move time for all jobs between machines is negligible.

7. There are no groups of similar machines.

8. Machines never break down.

9. Machines are never unable to perform a required task for lack of an operator, tool, or material.

10. The generation of the job route (usually by means of a transition matrix) does not allow consecutive operations on the same machine.

11. Alternate routing is not considered.

12. Lap-phasing is not permitted.

13. Lot-splitting is not considered.

14. No overtime or subcontracting is permitted (i.e., fixed shop capacity).

15. No allowance is made for scrap or rework.

16. Any statistics to be gathered are determined when the system is operating under steady state conditions.

Through any given article published in the area of dynamic job shops may not include all of the above assumptions, it will probably contain at least half of those listed.

Because of the stochastic nature of the parameters in dynamic job shops, Monte Carlo simulation has been the principle tool of analysis. A number of factors appear common to all these experimental simulations. Shop load parameters invariably include: (1) the mean arrival rate of jobs in the shop, (2) the mean processing rates at the various machines or machine centers, and (3) the number of machines or machine centers (including the number of machines in each center) in the shop (i.e., shop size). Operational characteristics of the system typically include: (1) the distribution of arrivals of jobs in the shop, (2) processing time distributions at the individual machines or machine centers, and (3) a procedure for generating job routings. And finally there is the queue discipline, the policy for sequencing jobs waiting in queue. Attention will focus on these factors in the following discussion.

Order arrivals. In many simulation studies of the dynamic job shop situation the familiar Poisson arrival rate has been assumed [30, 96, 103, 132, 139]. The mean for the Poisson distribution may be determined arbitrarily, but usual practice is to set it at a lower value than the mean service rate of the shop in order to prevent continual shop overload. Given that the times of arrival occur by a Poisson process, it can be shown that the interarrival times have a related exponential distribution [147, pg. 37]. Hence the exponential distribution is also representative of the Poisson process, but it describes the time between arrivals, and specifies that these time intervals are completely random.

Justification for using the Poisson distribution has been offered by Reinitz [138, pg. 60]. Empirical observations of job shops have revealed, according to Reinitz, that the arrivals of units to the various processing centers or departments approximate very closely the Poisson distribution. "But there is also a rational basis for using the Poisson process as a model for a real-system input. The Poisson process may be looked upon as a limit of disintegration of scheduled processes. The arrivals to a service department originate from completed services in other departments. Each process at each department has some distribution of service time, and the divergence of events from the scheduled times randomizes the input process, an intrinsic characteristic of the Poisson Law" [138, pg. 60].

Though much of the literature has assumed exponential interarrival times, a few investigations have used an Erlang distribution. Kuratani and Nelson [103] have concluded from empirical data that job arrivals in the shop actually follow a pattern "between" an exponential and an Erlang distribution.

Processing times. Processing time for a job at a given machine or machine center is generated from a service time distribution which may, for the sake of simulation economy, be assumed constant

for each machine (machine center) in the shop. However, one might well assign a different service time distribution to each machine (machine center). The exponential distribution has been used for generating service times in many dynamic job shop investigations [6, 27, 30, 96, 103, 139]. Other distributions suggested include the Erlang and hyper-exponential distribution [103], the normal distribution [26], and the lognormal distribution [1].

Harris [75] has questioned how well the Erlang distribution describes interarrival and service times in a real shop. Based on a study of a large machine shop in Los Angeles, he found that: (1) the observed queue lengths and waiting times are longer than predicted by an Erlang model, (2) the observed standard deviation in queue length is less than and the standard deviation in waiting time is greater than those predicted by an Erlang model, and (3) parametric changes in the Erlang model were not able to generate a single distribution which could predict what was in fact observed. Hence if one wants to simulate a real job shop, an investigation should be undertaken to determine the specific types of distributions peculiar to that shop.

Shop size. The third factor to be considered in the simulation of a dynamic job shop is the number of machines in the shop. If the simulation involves a real job shop, this number is readily available. However, if the simulation is synthetic, then a decision must be made concerning the number of machines to include in the model. Some of the smaller shop sizes that have been considered in synthetic simulations include five machines [1, 26, 29], six machines [27], and nine machines [6, 29, 30]. Real shop studies on the other hand have ranged from a 30-machine shop [42] to a shop composed of 1,000 machines [105].

In the literature on the synthetic simulation of dynamic job shops the number of machines comprising the shop have been small. However, Baker and Dzielinski [6], having tested shops of from 9 to 30 machines, found that the size of the shop was not significant. This result was in turn used by Conway, Johnson, and Maxwell [26] in justifying their simulation model of a five-machine shop. In a later article [27] Conway and Maxwell personally justified the use of a six-machine shop as adequate in representing the complexity involved in a large dynamic job shop situation. Nanot [127] has also found shop size to be an unimportant factor, "Since size of shop has never appeared as a major variable, it seems that we may be able to experiment with relatively small shops and generalize the resulting conclusions" [20, pg. 338].

Routing. Generation of job routings is the fourth factor to be considered in the simulation of dynamic job shops. Typically job routes are generated by means of a transition matrix, a stationary Markov chain. If there are N different machines in the shop, the Markov chain will consist of an N by $(N+1)$ matrix whose probabilities across any row of the matrix must sum to unity. The last column of the matrix represents the probability of leaving the shop as a completed order.

In order to use this transition matrix, the first machine in a job route must be specified. This may be done by random sampling from some predetermined distribution. Given the first machine assignment, the number of the next machine (if any) of the job route is chosen by generating a number from the distribution associated with the row of the Markov matrix which corresponds to the first machine assigned. If the number generated appears in the last column of the matrix, the job route consists of a single machine assignment. Otherwise, a second machine is assigned to the job route and this becomes the new row from which the third machine assignment will be generated. The process is repeated until the sequence is terminated by the selection of a number in the last column of the matrix.

If each row of the transition matrix contains equal probabilities the situation described is re-

ferred to as the "pure" job shop. A pure flow shop, on the other hand, contains only 0's or 1's as entries [27]. Both of these situations are undoubtedly rare in practice, but a great deal of the literature has dealt with these extreme cases [32, pg. 7].

In the simulation of a real dynamic job shop, the transition matrix must be determined from the data at hand. Kuratani and Nelson [103] describe two methods which might be considered in constructing this matrix. In the first method, termed the single operation dependence method, the transition probabilities are obtained by counting the number of times a machine (or machine center) is followed by each other machine (machine center) in the empirical data. For the second method, termed the double operation dependence method, the transition probabilities are determined by counting the number of times every permutation of two machines (machine centers) is followed by each machine (machine center) in the empirical data. Conceivably this could be extended to permutations of three machines, four machines, etc. At any rate, the experience and judgment of the investigator may determine how far this procedure should be carried.

In a synthetic simulation what effect might the transition matrix have on the performance of a given queue discipline? According to a study by Conway and Maxwell [27], the effect is not significant. In their study, both the pure job shop and pure flow shop transition matrices were studied. Surprising as it may seem, there was little difference in shop performance between the two cases.

As mentioned previously, many dynamic job shop simulations do not consider the possibility of of alternate routing. In a real shop simulation, however, alternate routing may be of great importance. Therefore, if the simulation model employs a transition matrix for generating a sequence of operations (the job route) for each job, alternate routing for a given job could be provided by randomizing the machine numbers in various segments of the job route. The assumption is that, within a given segment, there are no precedence relations to be considered. Hence any operation on a machine within a segment may precede any other in that segment. In this manner, a series of alternate routes may be generated. Studies [120, 146] have shown that shop performance can be improved if alternate routing is permitted in the model.

One final comment should be made with regard to the generation of job routes containing many more operations than there are machines in the shop. For example, if the shop contained only five machines, if would be possible to generate a job route of forty operations for an average of eight operations per machine. Many experimenters have eliminated such a possibility by truncating, at some predetermined point, the sequence generated from the transition matrix. Studies which have set an upper limit on the number of machines to be specified in a job route include: (1) Ackerman [1], who set an upper limit of seven machines, (2) Conway, Johnson and Maxwell [26], who also set an upper limit of seven, and (3) Kuratani and Nelson [103], who stopped the generation of the job route when more than 12 operations were generated.

Assignment of due dates. In the ideal situation, each job would be completed just at its due date. Management desires both predictability (because it would like to set due dates correctly) and control (because it wants to meet the due dates which are set). Completion of jobs ahead of the due date would result in storage costs if the products could not be shipped ahead of schedule. On the other hand, if jobs are completed later than the due date, customer dissatisfaction may result. Also, if any penalty clauses are included in the contract, they may be invoked by the customer.

Due date assignments are classified by Conway [30] into two types depending upon who is setting the due dates. First there are due dates which are exogenous and arbitrary—set by some external agency without regard to the processing characteristics of the shop itself, other jobs in the shop, or

the priority rule to be used. The problem is simply to find a procedure which is able to enforce such a set of due dates. Two such methods used by industry are what Conway terms the CON (constant) and RAN (random) methods [30]. For CON due dates the allowable shop time (i.e., the difference between the due date and the time the job arrives in the shop) is a constant amount, independent of any characteristic of the job to which it is assigned. This method represents the situation in which salesmen quote due dates uniformly on all orders. For example, they may promise delivery in five weeks from the time the order is placed. RAN due dates, in which the allowable shop time is assigned at random, correspond to situations in which the due date is chosen by the customer and accepted by the firm's salesmen.

In the second type of due date assignment, due dates are determined internally and are presumably rational. Two such methods mentioned by Conway [30] are what he has termed TWK (total work content) and NOP (number of operations). For TWK due dates, the allowable shop time is proportional to the sum of the processing times of the operations of a job. Errors in estimating these processing times apparently have little effect on overall shop performance if one is willing to extrapolate the conclusions of a study by Muth [121]. Allowable shop time for NOP due dates, on the other hand, is proportional to the number of operations. Both of these methods have been used industrially [64, pg. 562].

Types of priority rules. Although many priority rules have been proposed in the literature, no analytical formulation has been made which assures optimality [145]. Yet priority rules can do more than accelerate one or more jobs through the shop at the expense of other jobs—it can accelerate the average of all jobs [26]. Hence, though optimality will be difficult to achieve, it has been demonstrated that the choice of priority rule can affect aggregate measures of shop performance [27].

Conway, Johnson, and Maxwell [26] have classified priority rules into four categories:

1. Lateness rules—rules that determine priority according to some increasing function of lateness.
2. Arrival order rules—rules that assign priority according to the order in which jobs arrive at the machine under consideration.
3. Rules that determine priority according to some property of the job itself.
4. Random rule—a rule which assigns priorities at random.

The examples of priority rules given in the appendix are typical of those which have been investigated in the literature of dynamic job shop research. Conway [28] has investigated as many as 92 such rules though some of those rules were actually of the same form with different parameter settings. Nevertheless, it is the largest experimental investigation of priority rules reported to date.

The effectiveness of priority rules. Priority rules are compared on the basis of their effectiveness in meeting a predetermined criterion (or criteria). The effectiveness of the priority rule is measured by the value of the criterion function which results when the rule is followed. For a model of a specific waiting-line system, a measure of effectiveness would be an economic measure involving properly defined costs, revenues, and functional relationships based on a study of the system. For a general experimental model of waiting-line systems, however, one abstracts certain basic performance properties of the system which relate to possible measures of effectiveness for real systems, and then investigates the functioning of the system with respect to these basic properties using systematic experimentation based on a number of input and system parameters [132]. In order to eliminate one sizeable source of variability in the experimentation, each priority rule is tested on the same set of jobs, rather than drawing an independent sample of jobs for each experiment [26].

"Different industries have typically quite different goals. For some it is utilization of facilities (or labor). For others it is investment in work in progress, finished goods, raw material and total. They are

seriously concerned with service to the customer—completion of the work on or before the promised delivery date" [18]. In recognition of these concerns much of the experimental simulation work done in the area of dynamic job shops has been concerned with the effect of priority rules on the mean and variance of distributions describing the following measures of shop performance:

1. Job lateness with respect to due dates.
2. Percent of jobs late with respect to due dates.
3. Work-in-process inventory.
4. Order flow time.
5. Waiting time at each machine or for the shop as a whole.
6. Number of orders in the shop.
7. Machine utilization.
8. Setup time.

Attempts have also been made to study priority rules with respect to multiple criteria. One of the noteworthy studies in this area is reported by Legrande [105, pp. 151–153]. Six priority rules are examined with respect to 10 different weighted criteria.

A generalized statement of the relative effectiveness of priority rules in relation to a given criterion is difficult, if not impossible, to make. The reason becomes apparent when one considers the many possible types of distributions, transition matrices, parameter settings, and other assumptions which may be used in developing a simulation of a dynamic job shop. This is precisely the problem one encounters in studying the literature when noting, for example, that in one experiment priority rule Z was the most effective rule in relation to criterion Y, whereas in another experiment employing the same set of priority rules, priority rule Z was relatively ineffective.

Discussion of Popular Priority Rules

The shortest operation time rule. With this in mind, consider now the merits of the SOT rule (assume that setup time is included in the processing time) as applied to dynamic job shops. In a 1960 study by Baker and Dzielinski it was concluded that, if the average of the jobs' manufacturing times is used as a measure of effectiveness, the SOT rule is best [6]. Conway in a 1965 study compared the SOT rule to 31 other rules and found that, although the SOT rule did not exhibit the minimum value of any of the measures of work-in-process inventory studied, it nevertheless dominated the set of rules tested. According to Conway, the "performance under every measure was very good, it was an important factor in each of the rules that exhibit a 'best' performance under some measure, and it is simpler and easier to implement than the rules that surpass it in performance. It surely should be considered the 'standard' in scheduling research, against which candidate procedures must demonstrate their virtue"[29, pg. 129].

In a 1962 study [27], Conway and Maxwell tested the sensitivity of the SOT rule to errors in estimated processing time and to changes in the size of the shop. From their experiments, they concluded that the SOT rule was not sensitive to either of these two factors. Furthermore, they believed that, though a superior (nonlocal) rule could possibly be devised, it would have to be much more complex than the SOT rule.

In an extensive computer simulation experiment in 1965, Conway [30] investigated the performance of 39 priority rules in conjunction with the four methods of assigning due dates. For all the due-date methods tried, the SOT rule proved to be the least sensitive of the priority rules studied.

One of the biggest objections to the use of the SOT rule, however, is that while the mean time which jobs spend in the system is small, individual jobs (those with long operations) will be intolerably

delayed. Thus the variance of the lateness distribution is the basic disadvantage of the SOT rule. Three attempts have been reported by Conway [29, pg. 128] to reduce this variance:

1. Alternating the SOT rule with a low variance rule (with respect to flow time) to periodically clean out the shop.

2. Forcibly truncating the SOT rule by imposing a limit on the delay that individual jobs will tolerate.

3. Dividing jobs into two classes, preferred and regular. When the next job in the queue of the machine in question is to be chosen, the preferred job with the shortest operation time is selected. If there are no preferred jobs in this particular queue, then the regular job with the shortest processing time is selected.

Some success was achieved in reducing the variance of the lateness distribution with each of the above schemes without sacrificing the benefits of the SOT rule itself.

The COVERT rule. The formulation of the COVERT rule was another attempt to find a rule which retained the performance of the SOT rule, but which tended to minimize the extreme completion delays of a few orders [20, pg. 313]. The COVERT rule establishes, for each job in queue, a ratio whose numerator is the delay cost rate for the jth job and whose denominator is the processing time of that job on the machine in question. "The total delay cost is minimized if the task with the highest ratio is dispatched, since, in this process, one maximizes the tradeoff of the resource being allocated (processing time) against the potential gain (delay avoided)" [160, pg. 66].

The results obtained by Carroll [22] show the mean tardiness of the COVERT rule to be superior to even the truncated SOT rule. With respect to the six rules tested by Carroll, the lateness distribution was effectively skewed by the COVERT so that orders completed on the due date are the mode of the distribution with very few late orders. "It appears that the COVERT rule has a great deal to offer in maintaining the benefits of the SOT rule while dealing effectively with its defects" [20, pg. 317].

The slack time per remaining operation rule. In addition to the SOT rule, another rule frequently mentioned in the literature of dynamic job shops is the slack time per remaining operation rule. The careful reader of this literature will note that there are at least three ways of defining the term *slack*:

1. Static slack — due date less the time of arrival at the machine (or machine center) [20].

2. Dynamic slack.

   a. time remaining to due date less the sum of the remaining processing times and the remaining expected queue times [24].

   b. time remaining to due date less the remaining processing time [20, 105].

A comparison of the static slack per remaining operation rule and the first type of dynamic slack remaining operation rule can be found in [20]. The bases of comparison are the mean and variance of the flow time distributions for six different shop structures. Though one might expect the dynamic rule to yield better performance, the results of this experiment by Nanot do not justify this conclusion.

For the first type of dynamic slack per remaining operation rule, Colley asserts that, in comparison to other rules tested this rule will minimize both the variance of order completion intervals and the percentage of orders that are late [24]. Positive priority values assigned to jobs indicate that those jobs are ahead of schedule. A zero priority value indicates that a job is on schedule. However, if a job has a negative priority under this dynamic slack time per remaining operation rule, there is a possibility of that job missing its planned due date unless some action, such as expediting, is taken.

The second type of dynamic slack per remaining operation rule, listed above, has been studied by Conway [30]. Of the priority rules he considered under the TWK method of assigning due dates, this

type of dynamic slack per remaining operation rule yielded the lowest values for the variance of the completion time distribution and also for the percent of jobs completed before their assigned due dates. Since the SOT rule yielded the next lowest percentage of jobs late, Conway combined the two rules into a composite rule which gave better results than either of the two component rules considered separately [30].

Implementation of the second dynamic slack per remaining operation rule has been achieved in a large computer-monitored job shop. Bulkin, Colley, and Steinhoff [21] report that this priority rule determines the sequencing of jobs in the El Segundo Division of Hughes Aircraft Company, a job shop in which there are from 2,000 to 3,000 orders in progress at any one time. When the priority value of a job becomes negative under this rule, the job is immediately placed in an expediate status.

In a 1963 study reported by Legrande [105, pp. 146–151] this second type of dynamic slack per remaining operation rule was compared to five other rules including the SOT rule. The dynamic slack per remaining operation rule was found to rank first with respect to having the smallest standard deviation of the completion time distribution and the percent of orders completed late. SOT, on the other hand, ranked first with respect to earliest mean completion time, smallest average number of orders waiting in the shop, smallest average waiting time of orders in the shop, and largest percent of machine capacity utilized.

Priority rules employed in industry. Through discussing priority rules with many industrial people over a period of years, Conway [30] concluded that there are four rules in fairly common use where job lateness is of concern. These rules, listed in order of decreasing frequency of occurrence, are:

1. Assigning highest priority to jobs with the earliest due dates.

2. Assigning highest priority to the job with the least slack, where slack is defined as the time remaining until due date after deducting the remaining processing time.

3. Assigning highest priority to the job with the earliest due date at that machine. Allowable shop time is divided equally among the operations of the job to obtain the due date for each operation listed in the job route.

4. Assigning highest priority to the job which has the least slack per remaining operation. Slack is defined as in the second rule above.

However, with respect to general application, many dispatching rules used in industry come much closer to being classified as a longest operation time rule. "One reason for this is that an attempt to serve a criterion that is weighted by some index of job importance. People will often rank jobs by the index of importance along. Since it is reasonable to expect some positive correlation between importance and processing time this would tend to favor jobs with longer processing times" [27, pg. 71].

Simulation of Dynamic Job Shops

As previously mentioned, the primary method of studying dynamic job shop phenomena is through simulation models. The main components of these models have been discussed, namely, the order arrival distribution, the processing time distribution, shop size, routing, and the priority of dispatching rules. The role played by each of these factors in the simulation process will now be examined.

Figure 2 is a pictorial representation of a general simulation model [64, pg. 558]. The first step involves the generation of a job file, a reservoir of jobs to be subjected to the simulation process. An alternative approach might be to load jobs into the shop as soon as they arrive, in which case jobs are generated as the simulation proceeds.

Each job in the job file is assigned an arrival time, a routing of machines on which it will be proc-

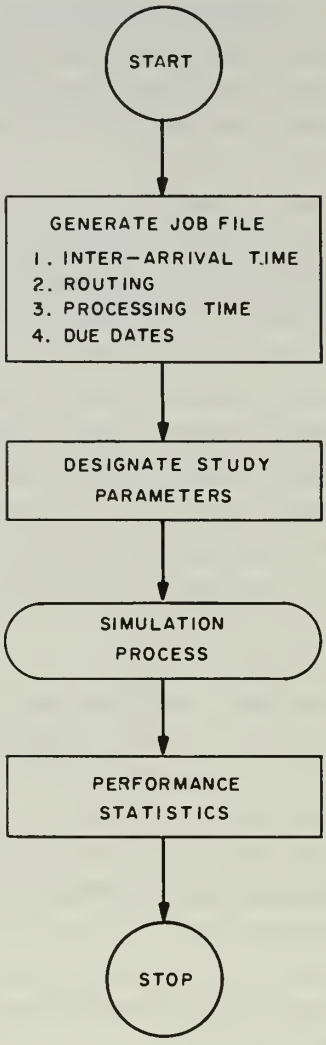


FIGURE 2. Pictorial representation of a general simulation model

essed, and the processing times associated with those machines. Setup times, processing costs, value of the job, or a due date might also be specified.

In addition to the parameters specified in the job file there are other parameters, termed study parameters, which are included in the simulation model. Such parameters might specify the number of times the computer will be interrogated to gather statistics on measures of performance, the number of machines in the shop, and the priority rules to be studied. Note that these parameters do not relate to any characteristic of a job itself.

Having specified the job file and the study parameters, the model is now prepared for the simulation process. Though the simulation could conceivably be handled by hand calculations, it more than likely will be performed on a computer. Each job is assigned to the shop on the basis of the interarrival time sequence. The computer clock is advanced until some event occurs such as the arrival of a job at the shop or the completion of an operation at a machine. At this time the computer will update the status of the shop. This updating might include the assignment of a newly arrived job to the queue of the first machine listed in its job route, the assignment of a job which has just completed an operation

to the next machine indicated by its job route, and the recording of accumulated job completion times, flow times, etc. The process continues, event by event, until the simulation is stopped. At this point the experimenter may choose to replicate the experiment using the same job file to provide a consistent basis for the comparison of performance statistics.

The output of the simulation includes various statistical measures of shop performance. One usually wishes that these statistics represent steady state conditions within the shop. At the start of the simulation the shop is completely idle. Work begins immediately on the first job which enters the shop. However, when the second job arrives, it may not be started as promptly as the first job, since the shop is still working on the first job. For the third, fourth, and all subsequent jobs, this type of delay occurs. After some period of time, however, these delays are stabilized, at which point steady state is said to have been achieved. The period of time necessary to achieve steady state is termed the warm-up interval [29]. Three other methods for initializing a shop simulation to eliminate the transient effects associated with the warm-up period have been suggested by Kuratani and Nelson [103].

Following the warm-up period, sampling of various measures of shop performance may begin. There are a number of ways to obtain a sample. The easiest, according to Conway, Johnson, and Maxwell [26], is to select a fixed number of consecutive jobs from the simulation output. Alternatively, one might designate, before simulation begins, a fixed number of consecutive jobs as the sample jobs and obtain the sample statistics when these jobs are discharged from the shop.

As with the transient effects which occur prior to steady state equilibrium, the experimenter should beware of this same type of phenomena when concluding a simulation run. If statistics are collected on the last jobs listed in the job file (shop run-out) the statistical results obtained will not be gathered under steady-state conditions. For example, the very last job discharged from the shop would not have had to compete as hard as its predecessors for machine capacity in the shop. In fact there were very few other orders in the shop while it was being processed. The phenomenon of shop run-out has been treated by Allen [4].

Various statistical methods have been mentioned in the literature with respect to the study of job shops. Correlation analysis [39], the Student's t test for the comparison of mean waiting times [145], regression analysis [35, 64, pg. 473] and analysis of variance [6, 35, 103] are a few examples. Factorial experiments have been conducted by Baker and Dzielinski [6], Conway and Maxwell [27], and Day [35].

Other Considerations in the Simulation of Dynamic Job Shops

In addition to interarrival times, processing times, routing, and due date, there are other parameters associated with a job which might be considered in the simulation of a dynamic job shop. These parameters include the value of the job, the costs involved in processing the job, setup time, the status of the job with respect to expediting, and lot sizes. Each of these will now be considered in order.

In 1960 Rowe [145] conducted a study in which the value of a job was made inversely proportional to its flow time allowance. The results appeared to show that decision rules which are based on the value of the jobs being processed provide reduced costs while still assuring a desired completion level. The simulation model used was based on an actual job shop containing 17 machine groups with a total of 85 machines.

What percentage of the jobs flowing through a shop have a relatively high value? Noting a 1951 study of value categories for a number of factories, Rowe [145, pg. 127] states that approximately 10 percent of the jobs in a factory constitute 75 percent of the total value of all jobs being processed. Conway, Johnson, and Maxwell [26] appear to acknowledge this finding in assigning jobs a dollar value from a distribution that was highly skewed to the right, specifically, the log normal distribution.

A second parameter which might be associated with each job is the cost involved in processing the job through the shop. Reinitz [138, pg. 72] separates this cost into three components:

1. Cost per unit of time of in-process inventory.
2. Set-up and tear-down cost.
3. Customer's delay cost rate.

Operationally, the customer's delay cost rate has the following meaning: "some contracts may specify penalties if delivery commitments are not met, and rewards for early deliveries. Such costs (rewards) may be of the form of dollars per unit time being late (early). In other cases the penalty (reward) of being late (early) may be associated with loss (gain) of anticipated profits. This can happen if late (early) deliveries cause loss (gain) of future business" [138, pg. 72].

In the preceding paragraph it will be noted that each of the component costs are functions of time. Manne [111] argues that if job shop management could reduce the time spent in the shop by all jobs, then, in effect, costs would also be reduced.

With respect to representing cost in the form of a distribution, Legrande [105, pg. 143] has used the negative exponential distribution. In his job shop simulation of the El Segundo Division of Hughes Aircraft Company, orders are assigned an initial cost, when the order is generated, by randomly sampling from a negative exponential distribution with a predetermined mean. The use of indifference contract curves for determining a lateness cost function has been suggested by Hershauer [80].

Set-up times are the third job parameter which might be considered in the simulation of a dynamic job shop. Elmaghraby and Cole [42] have treated set-up times in an integer programming context. However, most of the literature on dynamic job shops has included set-up time in the processing time of the job.

There is also the possibility of placing a job in an expedite status. Simulation studies of expediting in pure job shops have been reported by Hottenstein [84, 85] and these studies have been extended to hybrid job shops and to flow shops [86]. Given that a job is to be expedited, a decision must be made whether it will or will not be allowed to preempt other jobs at a given machine. If preemption is permitted, either a preempt repeat or preempt resume discipline must be assumed.

In much of the literature on job shop simulation, the problem of lot sizes has not been considered. However, both Reiter [139] and Beaird [9] have noted the incorporation of specific log size models in the simulation of job shop operations at Fairfield Manufacturing Company and General Electric, respectively. According to Beaird, an economic lot size model must deal with two opposing cost considerations. The first of these considerations is the fixed costs incurred when each lot is fabricated, specifically, machine set-up costs and paperwork costs. The opposing consideration is the variable cost of inventory handling, storage, obsolescence, and working capital tie-up incurred because of the premature processing of orders resulting from combining them with other more current orders in order to produce larger lot sizes.

Other Approaches to Studying the Dynamic Job Shop

As mentioned earlier, simulation has been the most frequently used method of analyzing the dynamic job shop situation. However, the simulation model may be improved by the inclusion of a set of heuristic rules to bolster the dispatching rules already in use. Gere [65] has successfully applied heuristic rules to the static $M \times N$ stochastic sequencing problem, and it appears that the inclusion of heuristics in the dynamic job shop model will also yield improved results. In fact, at the General Electric Company, heuristic programs have been incorporated into their computerized job shop planning and control system with great success [9].

Another approach to the study of dynamic job shops involves the scheduling behavior of people who are in charge of this function in the shop. Fox and Kriebel [60] have mathematically described certain aspects of the decision-making behavior of a manager in a New England shoe box manufacturing firm. In another study by Hurst and McNamara [87] the scheduling procedure of a production planner in a woolen mill was transformed into a mathematical decision model. A third study by Dutton [38] has specifically set out to simulate the behavior of a human scheduler who can analyze and classify orders into a few types.

Finally there is the dynamic programming approach to the study of the dynamic job shop simulation. From the survey of the literature, there has apparently been very little treatment of dynamic job shops using this approach. One article by Reinitz [138] develops a computer model for both overall shop activity and individual machine center operation; however, the model may prove unwieldy for smaller types of computers.

The Multi-Component and Machine-Labor Job Shop Situations

Thus far attention has focused mainly on the (1a–2a–3a) and (1a–2a–3b) situations. However, once these situations have been understood, research attention will probably focus on the more complex realistic job shops of the (1b–2b–3a) and (1b–2b–3b) variety.

To date, few, if any, studies have been made of the (1b–2b–3a) and (1b–2b–3b) situations. Examples of the studies which have been made of multi-component jobs and labor-machine situations are the following:

1. The (1b–2a–3a) situation [79, 81].
2. The (1b–2a–3b) situation [160].
3. The (1a–2b–3a) situation [4].
4. The (1a–2b–3b) situation [132].

SIMULATION AND COMPUTER SCHEDULING IN REAL SHOP SITUATIONS

For those job shops which can economically justify the rental or purchase of a computer, two possible approaches to monitoring shop operations are available—simulation or the installation of a real-time system. The essential difference in these two approaches is one of timeliness. With the simulation approach the computer is used daily or weekly to forecast possible outcomes of alternative sequencing strategies. Much of the computer input is randomly generated from functions which have been shown to approximately describe shop behavior. Given the current status of the shop, the randomly generated input is incorporated into a model which will generate alternative operating strategies for the coming period.

In real-time systems, on the other hand, the input to the computer comes directly from the shop office or shop floor when there is a change in the status of an order. In contrast to the simulation approach in which the computer clock is based on simulated time, the real-time system, as its name implies, operates in real time. Information in the real-time system is thus more timely and therefore more costly to maintain.

Effective use of job shop simulation, according to Legrande [105, pg. 136], will help shop management to do the following:

1. Establish more realistic scheduling wait times.
2. Establish more realistic order schedules.
3. Forecast shop load.

4. Plan equipment layout.

5. Test various sets of operating decisions.

If the simulation model is in complete agreement with the actual shop parameters, then the simulated and actual shop acitivty would be identical; however, due to the many assumptions built into the simulation model, this seldom occurs [21, pg. 186]. Hence one is faced with the question of validity, the degree to which the assumptions made in the simulation represent the situation in the real shop. An excellent treatment of this subject may be found in [44, pp. 437–440].

Literature articles describing the simulation of real job shops have been written by Allen [4], Bulkin, Colley, and Steinhoff [21], Colley [24 and 25], Elmaghraby and Cole [42], Legrande [105], Reiter [139], and Steinhoff [155]. Flow diagrams showing the relationships between the various variables and functions which may be incorporated into a simulation model of a real job shop can be found in Buffa [20, pp. 343–345], Legrande [105, pp. 138–141], Reiter [139] and Steinhoff [155, pg. 158].

Computer scheduling using a real-time system has been described by Beaird [9 and 10]. Four key preparation measures for the implementation of such a system and three measures for evaluating the effectiveness of the system are given [10]. A major portion of the system described has been implemented by the General Electric Company.

APPENDIX

Examples of Priority Rules

In order to discuss the calculation of job priority values (each of which is determined for the machine in question) the following notation is introduced:

a = a dummy variable.

$A_{j,k}$ = the arrival time of job j at the kth machine specified on its job route.

C_i = a constant ($i = 1, 2, \ldots, 12$).

D_j = the due date of job j.

E_j = the arrival time of job j in the shop.

i = a dummy subscript.

j = job number.

k = the kth machine specified in a given job's routing.

n_j = the nth (last) machine specified in the job route of job j.

$p_{j,k}$ = the processing time of job j on the kth machine specified in the job route of job j.

$\pi_{j,k}$ = the priority of job j at the kth machine specified in the job route of job j.

$s_{j,k}$ = the set-up time of job j on the kth machine specified in the job route of job j.

t = the time at which selection from a queue is to be made.

The examples are presented in outline form:

A. Lateness rules.

1. Inversely relating priority value to the due date of the job. The job with the earliest due date receives the highest priority [18, 26, 30, 105].

$$\pi_{j,k} = \frac{C_1}{D_j}.$$

2. Inversely relating priority value to the remaining slack time. Highest priority is given to the job for which the time remaining to due date less the remaining processing time is a minimum [26].

$$\pi_{j,k}=\frac{C_2}{D_j-\left[t+\sum_{i=k}^{nj} P_{j,i}\right]}.$$

3. Relating priority value directly to the remaining slack time. Highest priority is given to the job for which the time remaining to due date less the remaining processing time is a maximum [30].

$$\pi_{j,k}=D_j-\left[t+\sum_{i=k}^{nj} P_{j,i}\right].$$

4. Relating priority value directly to the ratio whose numerator is the slack time and whose denominator is the remaining processing time of the job. Highest priority is given to the job waiting in queue for which this ratio is smallest [103].

$$\pi_{j,k}=\frac{D_j-\left[t+\sum_{i=k}^{nj} P_{j,i}\right]}{\sum_{i=k}^{nj} P_{j,i}}.$$

5. Relating priority value directly to the slack time per remaining operation. Highest priority is assigned to the job with the smallest ratio of slack time to the number of operations remaining [18, 103, 105].

$$\pi_{j,k}=\frac{D_j-\left[t+\sum_{i=k}^{nj} P_{j,i}\right]}{nj-\sum_{a=1}^{k} 1}.$$

B. Arrival order rules.

1. Assigning priority on a first-come, first-served (FCFS) basis. The first arrival in the queue of the machine receives the highest priority: priority values are assigned to jobs in a strictly decreasing sequence as they arrive in the particular queue [18, 26, 28, 29, 30, 105, 145].

$$\pi_{j,k}=\frac{C_3}{A_{j,k}}.$$

2. Assigning priority on the basis of arrival time in the shop. Of the jobs waiting for service at a machine, highest priority would be assigned to the job with the earliest shop arrival date [29, 30].

$$\pi_{j,k}=\frac{C_4}{E_j}.$$

C. Rules that determine priority according to some property of the job itself.

1. Inversely relating priority value to the processing time of the coming operation. Highest priority is assigned to the job with the shortest processing time on the machine in question. This rule, as mentioned priviously, is synonymous with the SOT rule if set-up times are not considered [18, 26, 27, 28, 29, 30, 105, 145].

$$\pi_{j,k}=\frac{C_5}{P_{j,k}}.$$

2. Relating priority directly to the processing time on the coming operation. Highest priority is given to the job with the longest operation on the machine in question [26, 29, 30, 145].

$$\pi_{j,k}=C_6P_{j,k}.$$

3. Inversely relating priority value to the total remaining processing time. Highest priority is given to the job for which the sum of the processing times for all remaining operations is a minimum [26, 28, 29].

$$\pi_{j,k}=\frac{C_7}{\sum_{i=k}^{nj}P_{j,i}}.$$

4. Relating priority directly to the total remaining processing time. Highest priority is assigned to the job for which the sum of the processing time for all remaining operations is a maximum [26, 29].

$$\pi_{j,k}=C_8\sum_{i=k}^{ni}p_{j,i}.$$

5. Relating priority value directly to the total processing time of all operations on the job route. Highest priority is assigned to the job for which the total processing time of all operations on the job route is a maximum [28, 29].

$$\pi_{j,k}=C_9\sum_{i=1}^{ni}p_{j,i}.$$

6. Assigning priority on the basis of the ratio of processing time to the total work content of the job. Highest priority is assigned to the job which has the smallest ratio of processing time (on the machine in question) to the total processing time of all operations on the job route [29].

$$\pi_{j,k}=\frac{P_{j,k}}{\sum_{i=1}^{ni}P_{j,i}}.$$

7. Relating priority to the queue length of the next machine defined on the job route. Highest priority is given to the particular job which, upon leaving this machine, will go on to the next machine which has the shortest (in the sense of least processing time) critical queue. FCFS is used if no queue is critical. A queue is critical if it has less than a specified number of time units of processing time waiting. Actually, there is a family of rules of this type which can be parameterized by Q, a value below which a queue becomes critical [26].

8. Relating priority to the work content of the queue of the next machine defined on the job route. Highest priority is assigned to the particular job which will go on for its next operation to the queue with the least work, work being defined as the sum of the imminent processing times in that queue [28, 29].

9. Inversely relating priority value to the number of remaining operations. Highest priority is assigned to the job with the fewest remaining operations [26, 28, 29].

$$\pi_{j,k}=\frac{C_{10}}{n_j-\sum_{a=1}^{k}1}.$$

10. Relating priority directly to the number of remaining operations. Highest priority is assigned to the job with the most remaining operations [26].

$$\pi_{j,k}=C_{11}\left(n_j-\sum_{a=1}^{k}1\right).$$

11. Letting priority depend upon the (dollar) value of the job. Jobs are divided into high-value and low-value classes, greater priority being assigned to the high-value class of jobs. Within each class, the FCFS rule is observed. Actually, there is a family of rules of this type which can be parameterized by X, the proportion of jobs assigned to the high value class [26].

12. Relating priority value directly to the (dollar) value of the job. The priority is taken to be equal to the value of the job [26].

$$\pi_{j,k}=V_j.$$

13. Assigning priority according to the ratio of delay cost to processing time. Investigated in detail by Carroll [18], this family of priority rules is referred to as COVERT (cost over time) rules. Highest priority is given to the job for which the ratio is smallest. A complete example of the calculation of COVERT priority values is given in [20, pp. 312–315].

14. Assigning priority on the basis of the earliest operation due date. Equally spaced due dates are assigned to each operation on the job route at the time the job enters the shop. Highest priority is assigned to the job with the earliest operation due date [30].

$$\pi_{j,k}=A_{j,1}+(D_j-A_{j,1})\frac{K}{n_j}.$$

15. Assigning priority on the basis of set-up time. Highest priority is given to the job that requires the least amount of set-up time relative to the job which has just been completed on the machine in question [64].

$$\pi_{j,k}=\frac{C_{12}}{S_{j,k}}.$$

D. Random rule [27, 28, 29, 30, 105].

BIBLIOGRAPHY

[1] Ackerman, Sanford S., "Even-Flow, A Scheduling Method for Reducing Lateness in Job Shops," Management Technology Vol. 3, No. 1 (May 1963).

[2] Akers, S. B. and J. Friedman, "A Non-Numerical Approach to Production Scheduling Problems," Operations Research Vol. 3, No. 4 (Nov. 1955). Also discussed in Chapter 16 of Introduction to Operations Research, edited by C. W. Churchman, R. L. Ackoff, and E. L. Arnoff (John Wiley & Sons, Inc., New York, 1957), pp. 408–464.

[3] Akers, S. B., "A Graphical Approach to Production Scheduling Problems," Operations Research Vol. 4, No. 2 (Apr.–May 1956).

[4] Allen, Morton, "The Efficient Utilization of Labor Under Conditions of Fluctuating Demand," Industrial Scheduling edited by John F. Muth and Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[5] Avi-Itzhak, B., "A Sequence of Service Stations with Arbitrary Input and Regular Service Times," Management Science Vol. 11, No. 5 (Mar. 1965).

[6] Baker, C. T. and B. P. Dzielinski, "Simulation of a Simplified Job Shop," Management Science Vol. 6, No. 3 (1960), pp. 311–23.

[7] Balinski, M. L. and R. E. Quandt, "On an Integer Program for a Delivery Problem," Operations Research Vol. 12, No. 2 (Mar.–Apr. 1964).

[8] Banerjee, B. P., "Single Facility Sequencing with Random Execution Times," Operations Research Vol. 13, No. 3 (May–June 1965).

[9] Beaird, Samuel D., "A Practical Approach for Job Shop Planning and Control, Part 1," Automation Vol. 15, No. 3 (Mar. 1968).

[10] Beaird, Samuel D., "A Practical Approach for Job Shop Planning and Control, Part 2," Automation Vol. 15, No. 5 (May 1968).

[11] Beenhakker, H. L., "Development of Alternate Criteria for Optimality in the Machine Sequencing Problem," Ph.D. Thesis, Purdue University (1963).

[12] Bellman, R., "Mathematical Aspects of Scheduling Theory," Journal of SIAM *4*, 168–205 (1956).

[13] Bellman, R., "Dynamic Programming Treatment of the Traveling Salesman Problem," Journal of the Association for Computing Machinery Vol. 9, No. 1 (Jan. 1962).

[14] Bomberger, E. A., "A Dynamic Programming Approach to a Lot Size Scheduling Problem," Management Science Vol. 12, No. 11 (July 1966).

[15] Bowman, E. H., "The Schedule-Sequencing Problem," Operations Research Vol. 7, No. 5 (Sept. 1959).

[16] Brooks, George H. and Charles R. White, "An Algorithm for Finding Optimal or Near Optimal Solutions to the Production Scheduling Problem," Journal of Industrial Engineering Vol. 16 (1965).

[17] Brown, A. P. G. and Z. A. Lomnicki, "Some Application of the 'Branch-and-Bound' Algorithm to the Machine Scheduling Problem," Operational Research Quarterly Vol. 17, No. 2 (June 1966).

[18] Brown, Robert G., "Simulations to Explore Alternative Sequencing Rules," Naval Research Logistics Quarterly *15*, 281–286 (June 1968).

[19] Buffa, Elwood S., Models for Production and Operations Management (John Wiley and Sons, Inc., New York, 1963).

[20] Buffa, Elwood S., Production-Inventory Systems: Planning and Control (Richard D. Irwin, Inc., Homewood, Ill., 1968).

[21] Bulkin, Michael H., John L. Colley, and Harry W. Steinhoff, Jr., "Load Forecasting, Priority Sequencing, and Simulation in a Job-Shop Control System," in Readings in Production and Operations Management, by Elwood Buffa (John Wiley and Sons, Inc., New York, 1966).

[22] Carroll, D. C., "Heuristic Sequencing of Single and Multiple Component Jobs," Unpublished Ph. D. Dissertation, Sloan School of Management, M.I.T. (1965).

[23] Churchman, C. W., R. L. Ackoff, and E. L. Arnoff, Introduction to Operations Research (John Wiley and Sons, Inc., New York, 1957).

[24] Colley, John L., Jr., "Principles for Scheduling Job Shops," Systems and Procedures Journal Vol. 19, No. 4 (July–Aug. 1968).

[25] Colley, John L., Jr., "Implementing a Job-Shop Scheduling System," Systems and Procedures Journal Vol. 19, No. 5 (Sept.–Oct. 1968).

[26] Conway, Richard W., Bruce M. Johnson, and William L. Maxwell, "An Experimental Investigation of Priority Dispatching," Journal of Industrial Engineering Vol. XI, No. 3 (May–June 1960).

[27] Conway, R. W. and W. L. Maxwell, "Network Dispatching by the Shortest-Operation Discipline," Operations Research Vol. 10, No. 1 (Feb. 1962).

[28] Conway, R. W., "An Experimental Investigation of Priority Assignment in a Job Shop," Memorandum RM–3789–PR, The RAND Corporation, Santa Monica, California (Feb. 1964).

[29] Conway, Richard W., "Priority Dispatching and Work-In-Process Inventory in a Job Shop," Journal of Industrial Engineering Vol. 16, No. 2 (Mar.–Apr. 1965).

[30] Conway, Richard W., "Priority Dispatching and Job Lateness in a Job Shop," Journal of Industrial Engineering Vol. 16, No. 4 (July–Aug. 1965).

[31] Conway, R. W., W. L. Maxwell, and J. W. Oldziey, "Sequencing Against Due-Dates," Proceedings of the 4th IFORS (John Wiley and Sons, Inc., New York, 1966).

[32] Conway, Richard W., William L. Maxwell, and Louis W. Miller, Theory of Scheduling (Addison-Wesley Publishing Company, Reading, Massachusetts, 1967).

[33] Dantzig, G., D. R. Fulkerson, and S. M. Johnson, "On a Linear Programming Combinatorial Approach to the Traveling Salesman Problem," Operations Research Vol. 7, No. 1 (Jan.–Feb. 1959).

[34] Dantzig, G., "A Machine-Job Scheduling Problem," Management Science Vol. 6, No. 2 (Jan. 1960).

[35] Day, James E., "The Effect of Advancing Previously Established Due Dates on Various Performance Measures of a Pure Job Shop," Ph. D. Thesis, Pennsylvania State University (1970).

[36] Dudek, Richard A. and Ottis Foy Teuton, Jr., "Development of M-Stage Decision Rule for Scheduling N Jobs Through M Machines," Operations Research Vol. 21, No. 3 (May–June 1964).

[37] Dudek, R. A. and P. M. Ghare, "Make-Span Sequencing on M-Machines," The Journal of Industrial Engineering Vol. XVIII, No. 1 (Jan. 1967).

[38] Dutton, J. M., "Production Scheduling—A Behavioural Model," International Journal of Production Research Vol. 3, No. 3 (1964).

[39] Dzielinski, B. P., C. T. Baker, and A. S. Manne, "Simulation Tests of Lot-Size Programming," in Industrial Scheduling, edited by John F. Muth and Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1963).

[40] Eastman, W. L., S. Ever, and I. M. Isaacs, "Bounds for Optimal Scheduling of Jobs," Management Science Vol. 11, No. 2 (Dec. 1964).

[41] Eilon, S., "Economic Batch Size Determination of Multi-Product Scheduling," Operational Reserach Quarterly Vol. 9, No. 4 (Dec. 1958).

[42] Elmaghraby, Salah E. and Roger T. Cole, "On the Control of Production in Small Job Shops," Journal of Industrial Engineering Vol. 14, No. 4 (July–Aug. 1963).

[43] Elmaghraby, Salah E. and Allen S. Ginsberg, "A Dynamic Model for the Optimal Loading of Linear Multi-Operation Shops," Management Technology Vol. 4, No. 1 (June 1964).

[44] Elmaghraby, Salah E., The Design of Production Systems (Reinhold Publishing Company, New York, 1966).

[45] Elmaghraby, S. E., "The Minimum Number of Machines to Accomplish N Immediately Available Tasks," Research Memorandum, North Carolina State University (July 1967).

[46] Elmaghraby, S. E., "The Sequencing of N Jobs on M Parallel Processors," Research Memorandum, North Carolina State University (Jan. 1968).

[47] Elmaghraby, Salah E., "The One Machine Sequencing Problem With Delay Costs," The Journal of Industrial Engineering Vol. 19, No. 2 (Feb. 1968).

[48] Elmaghraby, Salah E., "The Sequencing of 'Related' Jobs," Naval Research Logistics Quarterly *15*, 23–32 (Mar. 1968).

[49] Elmaghraby, Salah E., "The Machine Sequencing Problem-Review and Extensions," Naval Research Logistics Quarterly Vol. *15*, (June 1968).

[50] Emery, James C., "An Approach to Job Shop Scheduling Using a Large-Scale Computer," Industrial Management Review Vol. 3, No. 1 (Fall, 1961).

[51] Emmons, Hamilton, "One Machine Sequencing to Minimize Total Tardiness," Dept. of Operations Research, Cornell University Tech. Report No. 52 (Sept. 1968).

[52] Evans, R. V., "Queueing When Jobs Require Several Services Which Need Not Be Sequenced," Western Management Science Institute, UCLA, Working Paper No. 17 (Sept. 1962).

[53] Evans, R. V., "Numerical Methods for Queues with Complex Service Rates," Western Management Science Institute, UCLA, Working Paper No. 29 (1963).

[54] Evans, R. V., "The Structure of Some Two Dimensional Queueing Systems," Western Management Science Institute, UCLA, Working Paper No. 36 (July 1963).

[55] Evans, R. V., "The Structure of Production and Processing Systems," Western Management Science Institute, UCLA, Working Paper No. 56 (Dec. 1964).

[56] Fabrycky, W. J. and J. E. Shamblin, "A Probability Based Sequencing Algorithm," The Journal of Industrial Engineering Vol. XVII, No. 6 (June 1966).

[57] Fife, Dennis W., "Scheduling with Random Arrivals and Linear Cost Functions," Management Science Vol. 11, No. 3 (Jan. 1965).

[58] Fisher, H. and G. L. Thompson, "Probabilistic Learning Combinations of Local Job-Shop Scheduling Rules," in Industrial Scheduling, edited by John F. Muth and Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[59] Ford, D. R., and L. R. Fulkerson, Flows in Networks (Princeton University Press, Princeton, New Jersey, 1962).

[60] Fox, P. D. and C. H. Kriebel, "An Empirical Study of Scheduling Decision Behavior," The Journal of Industrial Engineering (June 1967).

[61] Friedman, H. D., "Reduction Methods for Tandem Queueing Systems," Operations Research Vol. 13, No. 1 (Jan. 1965).

[62] Gapp, W., P. S. Mannkekar, and L. G. Mitten, "Sequencing Operations to Minimize In-Process

Inventory Costs," Management Science, Vol. 11, No. 3 (Jan. 1965).

[63] Gavett, J. William, "Three Heuristic Rules for Sequencing Jobs to a Single Production Facility," Management Science, Vol. 11, No. 8 (June 1965).

[64] Gavett, J. William, Production and Operations Management (Harcourt, Brace, and World, Inc., New York, 1968).

[65] Gere, William S., "Heuristics In Job Shop Scheduling," Management Science Vol. 13, No. 13 (Nov. 1966).

[66] Giffler, B. and G. L. Thompson, "Algorithms for Solving Production-Scheduling Problems," Operations Research Vol. 8, No. 4 (Oct.–Dec. 1960).

[67] Giffler, Bernard. "Schedule Algebras and Their Use in Formulating General Systems Simulations," in Industrial Scheduling, edited by John F. Muth & Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[68] Giffler, B., G. L. Thompson, and V. Van Van Ness, "Numerical Experience with Linear and Monte Carlo Algorithms for Solving Production Scheduling Problems," in Industrial Scheduling, edited by John F. Muth & Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[69] Giglio, R. J. and H. M. Wagner, "Approximate Solutions to the Three-Machine Scheduling Problems," Operations Research Vol. 12, No. 2 (Mar. 1964).

[70] Gilmore, P. and R. E. Gomory, "Sequencing of One State-Variable Machine," Operations Research Vol. 12, No. 5 (Sept.–Oct. 1964).

[71] Glassey, C. R., "Scheduling Several Products on One Machine to Minimize Changeovers," Operations Research Vol. 16, No. 2 (Mar.–Apr. 1968).

[72] Glassey, C. R., "Minimum Change-over Scheduling of Several Products on One Machine," Operations Research Vol. 16, No. 2 (Mar.–Apr. 1968).

[73] Greensberg, Harold, "A Branch-Bound Solution to the General Scheduling Problem," Operations Research Vol. 16, No. 2 (Mar.–Apr. 1968).

[74] Hardgrave, W. W. and G. Nemhauser, "A Geometric Model and a Graphical Algorithm for a Sequencing Problem," Operations Research Vol. 11, No. 6 (Nov.–Dec. 1963).

[75] Harris, R. D., "An Empirical Investigation and Model Proposal of a Job Shop-Like Queueing System," Western Management Science Institute, UCLA, Working Paper No. 84 (July 1965).

[76] Held, Michael, Richard M. Karp, and Richard Sherashian, "Scheduling with Arbitrary Profit Functions," IBM Data Systems Division, Math & Appl. Dept. M & A–11 (1961).

[77] Held, M. and R. H. Karp, "A Dynamic Programming Approach to Sequencing Problems," Journal of SIAM Vol. 10, No. 1 (Mar. 1962).

[78] Heller, J., "Some Numerical Experiments for an M × J Flow Shop and Its Decision-Theoretical Aspects," Operations Research Vol. 8, No. 2 (Mar.–Apr. 1960).

[79] Heller, Jack and George Logemann, "An Algorithm for the Construction and Evaluation of Feasible Schedules," Management Science Vol. 8, No. 2 (Jan. 1962).

[80] Hershauer, James, "An Empirically-Derived Simulation to Explore Sequencing Decisions," Ph.D. Thesis, Indiana University (1969).

[81] Heuser, William A., Jr., and B. E. Wynne, Jr., "An Application of the Critical Path Method to Job Shop Scheduling—A Case Study," Management Technology, Vol. 3, No. 2 (Dec. 1963).

[82] Hillier, Frederick S., "Cost Models for the Application of Priority Waiting-Line Theory to Industrial Problems," Chapter 12 in Readings in Production and Operations Management, by

Elwood S. Buffa (John Wiley & Sons, Inc., New York, 1966).

[83] Holt, Charles C., "Priority Rules for Minimizing the Cost of Queues in Machine Scheduling," in Industrial Scheduling, edited by John F. Muth & Gerald P. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[84] Hottenstein, Michael P., "The Effects of Expediting On Measures of Performance In a Job Shop," A research report sponsored by the American Production and Inventory Control Society, University Park, Pennsylvania (Oct. 1968).

[85] Hottenstein, Michael P., "A Simulation Study of Expediting in a Job Shop," Journal of Production and Inventory Management Vol. 10, No. 2 (Second Quarter, 1969).

[86] Hottenstein, Michael P., "Expediting in Job-Order-Control Systems: A Simulation Study," AIIE Transactions (forthcoming).

[87] Hurst, E. Gerald, Jr. and Aidan B. McNamara, "Heuristic Scheduling In a Woolen Mill," Management Science Vol. 14, No. 4 (Dec. 1967).

[88] Ignall, Edward and Linus Schrage, "Application of the Branch and Bound Technique to Some Flow-Shop Scheduling Problems," Operations Research Vol. 13, No. 3 (May–June 1965).

[89] Jackson, James R., "An Extension of Johnson's Results on Job Lot Scheduling," Naval Research Logistics Quarterly *3*, 201–203 (Sept. 1956).

[90] Jackson, James R., "Networks of Waiting Lines," Operations Research Vol. 5 (Aug. 1967).

[91] Jackson, James R., "Simulation Research on Job Shop Production," Naval Research Logistics Quarterly *4*, 287–295 (Dec. 1957).

[92] Jackson, J. R., "Some Problems of Queueing with Dynamic Priorities," Naval Research Logistics Quarterly 7 (Sept. 1960).

[93] Jackson, James R., Rosser T. Nelson, and Andrew A. Grindlay, "Research on Jobshop-Like Models: A Progress Report," Management Sciences Research Project, Research Report No. 76, UCLA (July 18, 1962).

[94] Jackson, James R., "Networks of Queues with Controlled Service Rates," Management Sciences Research Project, Research Report No. 77 (August 28, 1962).

[95] Jackson, James R., "Queues with Dynamic Priority Discipline," in Industrial Scheduling, edited by John F. Muth and Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[96] Jackson, James R., "Jobshop-Like Queueing Systems," Management Science Vol. 10, No. 1 (Oct. 1963).

[97] Johnson, S. M., "Optimal Two- and Three-Stage Production Schedules with Setup Times Included," Naval Research Logistics Quarterly *1*, 61–68 (1954).

[98] Johnson, S. M., "Discussion: Sequencing N Jobs on 2 Machines with Arbitrary Time Lags," Management Science Vol. 5, No. 3 (Apr. 1959).

[99] Karush, W. and L. A. Moody, "Determination of Feasible Shipping Schedules for a Job Shop," Operations Research Vol. 6, No. 1 (Jan.–Feb. 1958).

[100] Karush, William, "A Counterexample to a Proposed Algorithm for Optimal Sequencing of Jobs," Operations Research Vol. 13, No. 2 (Mar. 1965).

[101] Kingman, J. F. C., "Two Similar Queues in Parallel," Annals of Mathematical Statistics, Vol. 32, No. 4 (Dec. 1961).

[102] Kortanek, K. O. and W. L. Maxwell, "On a New Class of Combination Optimizers for Multi-Product Single-Machine Scheduling," Cornell University, Dept. of Ind. Eng. and Operations Research, Technical Report No. 25 (June 1967).

[103] Kuratani, Yoshiro and Ross T. Nelson, "A Pre-Computational Report on Job-Shop Simulation Research," Journal of the Operations Research Society of Japan Vol. 2, No. 4 (Mar. 1960).

[104] Lawler, Eugene L., "On Scheduling Problems with Deferral Costs," Management Science Vol. 11, No. 2 (Nov. 1964).

[105] LeGrande, Earl, "The Development of a Factory Simulation System Using Actual Operating Data," in Readings in Production and Operations Management, by Elwood S. Buffa (John Wiley & Sons, Inc., New York, 1966).

[106] Lightenberg, E., "Minimal Cost Sequencing of N Grouped and Ordered Jobs on M Machines," Journal of Industrial Engineering Vol. 17, No. 4 (Apr. 1963).

[107] Little, J. D. C., K. G. Murty, D. W. Sweeny, and C. Karel, "An Algorithm for the Traveling Salesman Problem," Operations Research Vol. 11, No. 6 (Nov.–Dec. 1966).

[108] Lomnicki, Z. A., "A 'Branch-and-Bound' Algorithm for the Exact Solution of the Three-Machine Scheduling Problem," Operational Research Quarterly Vol. 16, No. 1 (Mar. 1965).

[109] Madigan, J. G., "Scheduling a Multi-Product Single Machine System for an Infinite Planning Period," Management Science Vol. 14, No. 11 (July 1968).

[110] Makino, Toji, "On a Scheduling Problem," Journal of the Operations Research Society of Japan Vol. 8, No. 1 (Sept. 1965).

[111] Manne, Alan S., "On the Job-Shop Scheduling Problem," Operations Research Vol. 8, No. 2 (Mar.–Apr. 1960).

[112] Maxwell, William L., "The Scheduling of Single Machine Systems—A Review," Dept. of Ind. and Eng. Adm., Cornell University (May 19, 1962).

[113] Maxwell, William L., "The Scheduling of Economic Lot Sizes," Naval Research Logistics Quarterly *11*, 89–124 (June–Sept. 1964).

[114] McMahan, G. B. and P. G. Burton, "Flow-Shop Scheduling with the Branch-and-Bound Method," Operations Research Vol. 15, No. 3 (May–June 1967).

[115] McNaughton, Robert, "Scheduling with Deadlines and Loss Functions," Management Science Vol. 6, No. 1 (Oct. 1959).

[116] Mellor, P., "A Review of Job Shop Scheduling," Operational Research Quarterly Vol. 17, No. 2 (June 1966).

[117] Mitten, L. G., "A Scheduling Problem," The Journal of Industrial Engineering Vol. X, No. 2 (Mar.–Apr. 1959).

[118] Mitten, L. G., "Sequencing N Jobs on Two Machines with Arbitrary Time Lags," Management Science Vol. 5, No. 3 (Apr. 1959).

[119] Moore, J. Michael, "An N Job One Machine Sequencing Algorithm for Minimizing the Number of Late Jobs," Management Science Vol. 15, No. 1 (Sept. 1968).

[120] Morse, P. M., Queues, Inventories, and Maintenance (John Wiley and Sons, Inc., New York, 1958).

[121] Muth, John F. "The Effect of Uncertainty in Job Times on Optimal Schedules," in Industrial Scheduling, edited by John F. Muth and Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[122] Nabeshima, Ichiro, "The Order of n Items Processed on m Machines," Journal of the Operations Research Society of Japan Vol. 3, No. 4 (Mar. 1961).

[123] Nabeshima, Ichiro, "The Order of n Items Processed on m Machines (II)," Journal of the Operations Research Society of Japan Vol. 4, No. 1 (Sept. 1961).

[124] Nabeshima, Ichiro, "Sequencing on Two Machines with Start Lag and Stop Lag," Journal of the Operations Research Society of Japan Vol. 5, No. 3 (June 1963).

[125] Nabeshima, Ichiro, "Computational Solution to the m Machine Scheduling Problem," Journal of the Operations Research Society of Japan Vol. 7, No. 3 (Feb. 1965).

[126] Nabeshima, Ichiro, "On the Bound of Makespans and Its Application in M Machine Scheduling Problem," Journal of the Operations Research Society of Japan Vol. 9, Nos. 3 & 4 (July 1967).

[127] Nanot, Y. R., "An Experimental Investigation and Comparative Evaluation of Priority Disciplines in Job Shop-Like Queueing Networks," Management Sciences Research Project, UCLA, Research Report No. 87 (1963).

[128] Nelson, R. T., "Waiting-Time Distributions for Application to a Series of Service Centers," Operations Research Vol. 6 (Nov.–Dec. 1958).

[129] Nelson, R. T., "An Empirical Study of Arrival, Service Time and Waiting Time Distributions of a Job Shop Production Process," Management Sciences Research Project, UCLA, Research Report No. 60 (June 1959).

[130] Nelson, R. T., "A Simulation Study and Analysis of a Two Station Waiting Line-Network Model," Management Sciences Research Project, UCLA, Research Report No. 91 (Jan. 1965).

[131] Nelson, R. T., "Dual Resource Constrained Series Service Systems," Western Management Science Institute, UCLA, Working Paper No. 113 (Feb. 1967).

[132] Nelson, Rosser T., "Labor and Machine Limited Production Systems," Management Science Vol. 13, No. 9 (May 1967).

[133] Nugent, C. E., "On Sampling Approaches to the Solution of the n by m Static Sequencing Problem," Ph.D. Thesis, Cornell University (Sept. 1964).

[134] Oldfather, Paula M., Allen S. Ginsberg, Patricia L. Love, and Harry M. Markowitz, "Programming by Questionnaire: The Job Shop Simulation Program Generator," RAND Memo, RM–5162–PR (July 1967).

[135] Palmer, D. S., "Sequencing Jobs Through a Multi-Stage Process in the Minimum Total Times," Operational Research Quarterly Vol. 16, No. 1 (1965).

[136] Presby, J. T. and M. L. Wolfson, "An Algorithm for Solving Job Sequencing Problems," Management Science Vol. 13, No. 8 (April 1967).

[137] Reich, E., "Waiting Times When Queues Are in Tandem," Annals of Mathematical Statistics Vol. 28, No. 3 (1957).

[138] Reinitz, Rudolph, "On the Job Shop Scheduling Problem," in Industrial Scheduling, edited by John F. Muth & Gerald P. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[139] Reiter, Stanley, "A System for Managing Job-Shop Production," The Journal of Business Vol. XXXIX, No. 3 (July 1966).

[140] Rogers, J., "A Computational Approach to the Lot Scheduling Problem," Management Science, Vol. 4, No. 3 (April 1958).

[141] Root, James G., "Scheduling with Deadlines and Loss Functions on k Parallel Machines," Management Science, Vol. 11, No. 3 (1965).

[142] Rothkopf, Michael, "Scheduling Independent Tasks on Parallel Processors," Management Science Vol. 12, No. 5 (Jan. 1966).

[143] Rothkopf, Michael, "Scheduling with Random Service Times," Management Science Vol. 12, No. 9 (May 1966).

[144] Rowe, Alan J., "Job Shop Scheduling Decision Rules—An Example of Computer Simulation,"

Managing Progress through Operations Research, Proceedings of 3rd Annual Conference, Society for the Advancement of Management, Hotel New Yorker, New York City (Feb. 6–7, 1958).

[145] Rowe, Alan J., "Toward a Theory of Scheduling," The Journal of Industrial Engineering Vol. XI, No. 2 (Mar.–Apr. 1960).

[146] Russo, F. J., "A Heuristic Approach to Alternate Routing in a Job Shop," Master's Thesis, MIT (June 1965).

[147] Saaty, Thomas L., Elements of Queueing Theory (McGraw-Hill Book Company, Inc., New York, 1961).

[148] Sandeman, P., "Empirical Design of Priority Waiting Times for Jobbing Shop Control," Operations Research Vol. 9, No. 4 (July–Aug. 1961).

[149] Schrage, Linus E. and Louis W. Miller, "The Queue M/G/I with the Shortest Remaining Processing Time Discipline," Operations Research Vol. 14, No. 4 (July–Aug. 1966).

[150] Sisson, Roger L., "Methods of Sequencing in Job Shops—A Review," Operations Research Vol. 7, No. 1 (Jan.–Feb. 1959).

[151] Sisson, R. L., "Sequencing Theory," in Progress in Operations Research, edited by R. M. Ackoff (John Wiley & Sons, Inc., New York, 1961), Chap. 7.

[152] Smith, Richard D. and Richard A. Dudek, "A General Algorithm for Solution of the n-Job, M-Machine Sequencing Problem of the Flow Shop," Operations Research (Dec. 1966).

[153] Smith, Wayne E., "Various Optimizers for Single-Stage Production," Naval Research Logistics Quarterly *3*, 59–66 (Mar.–June 1956).

[154] Spinner, Allen H., "Sequencing Theory-Development to Date," Naval Research Logistics Quarterly *15*, 319–330 (June 1968).

[155] Steinhoff, Harry W., Jr., "Daily System for Sequencing Orders in a Large-Scale Job Shop," Readings in Production and Operations Management, by Elwood S. Buffa (John Wiley and Sons, Inc., New York, 1966).

[156] Story, Alfred E. and Harvey M. Wagner, "Computational Experience with Integer Programming for Job-Shop Scheduling," Industrial Scheduling, edited by John F. Muth and Gerald L. Thompson (Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1963).

[157] Szwarc, Wlodzimierz, "Solution of the Akers-Friedman Problem," Operations Research Vol. 8, No. 6 (Nov.–Dec. 1960).

[158] Talwar, P. P., "A Note on Sequencing Problem with Uncertain Job Times," Journal of the Operations Research Society of Japan Vol. 9, Nos. 3 and 4 (July 1967).

[159] Thompson, G. L., "Recent Developments in the Job-Shop Scheduling Problem," Naval Research Logistics Quarterly 7, 585–589 (Dec. 1960).

[160] Trilling, Donald R., "Job Shop Simulation of Orders That Are Networks," Journal of Industrial Engineering Vol. XVII, No. 2 (Feb. 1966).

[161] Wagner, Harvey M., "An Integer Linear-Programming Model for Machine Scheduling," Naval Research Logistics Quarterly *6*, 131–140 (June 1959).

[162] Wilkins, C. A., "On Two Queues in Parallel," Biometrika *47*, 198 (1960).

* * *

COMPUTER-AIDED SELECTION OF PRIOR DISTRIBUTIONS FOR GENERATING MONTE CARLO CONFIDENCE BOUNDS ON SYSTEM RELIABILITY*

Nancy R. Mann

Rocketdyne
A Division of North American Rockwell Corporation
Canoga Park, California

ABSTRACT

A description is given of results of preliminary investigations (by a group at North American Rockwell Corporation) related to the Monte Carlo generation of lower confidence bounds on the reliability of a logically complex system. In calculating system confidence bounds by use of a Monte Carlo procedure, one must generate the distribution of each independent subsystem reliability, given the life-test failure data for that subsystem. Therefore, an assumption of a specified a priori distribution for each subsystem reliability is implicit in the procedure.

In order that clues may be obtained as to optimum prior assumptions to be used in calculating Monte Carlo bounds for a complex system, the model has been restricted to a series system wherein each independent subsystem has exponentially distributed failure time and prototypes of each subsystem are tested until a fixed (but not necessarily the same for each subsystem) number of failures occurs. For this model, optimum (uniformly most accurate unbiased) exact classical confidence bounds on the reliability $R(t_m)$ at a specified mission time, t_m, are available, although not easily calculated (El Mawaziny [12] and Lenter and Buehler [27]). Computer programs for calculating the optimum classical bounds and the Bayesian Monte Carlo bounds were written, and a means of numerically comparing various forms of prior distributions against an optimum standard was thus provided. One prior distribution widely used in obtaining Monte Carlo and general Bayesian exact lower confidence bounds on system reliability is thereby shown numerically to yield bounds which are conservative in the classical sense for this series-system model. Another suggested prior distribution is shown to give bounds which are usually conservative but under certain conditions are liberal, and hence not truly confidence bounds. Moreover, it is demonstrated by a combination of numerical and analytical results, that for a series system containing more than one independent subsystem there exists no prior distribution for subsystem reliability which is independent of the data and which yields the optimum lower bounds. Other numerical results related to the selection of optimum methods for generating the bounds and evaluation of certain approximate methods are described.

BACKGROUND AND APPROACH

Review of Pertinent Literature

If it is possible to determine confidence bounds on system reliability solely from the testing of the subsystems of which the system is comprised, saving of expensive system testing can be effected. It may, in fact sometimes be infeasible to test the system as a whole. Furthermore, this method of obtaining system confidence bounds can be used for exploratory system design.

The subject of confidence bounds for system reliability from subsystem testing is one about which much has been written, but not a great deal is known. Consider a series system in which the failure

*This research was sponsored by the Mathematics and Statistics Panel of the Aerospace and Systems Group (A&SG) of North American Rockwell Corporation and funded under the Internal Research and Development program of the Executive Offices of A&SG.

times of k independent subsystems are exponentially distributed, i.e., for T a random variable representing failure time, Prob $(T > t \equiv R(t) = \exp(-\lambda t)$, $t \geq 0$, $\lambda > 0$. Suppose n_j prototypes of the jth subsystem, are subjected to life test and the life test is terminated at the time of the r_jth ordered failure, $j = 1, 2, \ldots, k$ For this special model, there exist optimum (uniformly most accurate unbiased)* exact† confidence bounds on the reliability, $R(t_m)$, at time, t_m, the probability that the system will survive at least until time, t_m. (See Lentner and Buehler [27] and El Mawaziny [12]). No such optimum bounds have been found for a model which is equivalent to this exponential-failure-time series-system model, except for the fact that total test time, t_j, rather than number of failures, r_j, is specified for the life test of jth subsystem, $j = 1, 2, \ldots, k$, and number of failures is the observable random variable. For either the fixed-time or fixed-number-of-failures model, optimum exact confidence bounds have not been derived for cases in which either failure time has other than an exponential distribution (or can be converted by a transformation of the data to an exponential distribution) or the system is other than an independent *series* system.

Another much used failure model, often called the "attribute" model, is one in which only pass-fail binomially distributed data are collected for each independent subsystem. For this model, optimum exact confidence bounds on reliability (or probability of successful operation) of a series system have been derived (see Buehler [6]), but the problem of actually constructing such optimum bounds has not been completely solved (see Lipow [31, 32], Lloyd and Lipow [33], Schick [43], and Steck [48]. If a Poisson approximation to the binomial distribution is applicable, then results of Harris [22] provide optimal exact bounds on the reliability of an independent series system for the attribute model, if one randomizes appropriately in obtaining the bounds. One would expect the Poisson approximation to the binomial distribution to apply when the number of prototypes of each subsystem tested is large and the probability of failure for each subsystem is small. There appears to be some question, however (see Garner [19]), as to whether the approximation loses its applicability as the number of subsystems increases.

Many approximate and nonoptimal exact confidence bounds on system reliability have been derived. There have been several approximate confidence bounds on system reliability at time, t_m, derived for the exponential fixed-number-of-failures model wherein the independent subsystems form a series system. Some of the papers containing these derivations were written prior to the publication of the derivation of the optimum bounds (see Takenaga [49] and Kraemer [24]).

Other work has been directed at providing a more tractable method of calculating confidence bounds than that of El Mawaziny's generalization to k subsystems, $k \geq 2$, of the Lentner-Buehler bounds which apply to two subsystems only (see El Mawaziny and Buehler [13], Grubbs [21], and Sarkar [41]). The method suggested by El Mawaziny and Buehler depends upon large-sample theory and the others use the fact that a function of the estimator of subsystem mean-time-to-failure has a chi-square distribution. The method of Sarkar does not require that the subsystems be independent and is exact for equal numbers of failures for all subsystems.

Some rather limited numerical comparisons have been made of some of these nonoptimal methods for obtaining confidence bounds by, for example, Grubbs [21] and Sarkar [41]. Apparently none of these methods have until this time been subjected to a thorough comparison with the Lentner-Buehler-

*The definitions of uniformly most accurate and unbiased confidence bounds are as given by Lehman [26]. They are as follows: A confidence bound $\underline{\theta}(X)$ satisfying $P_\theta\{\underline{\theta}(X) \leq \theta\} \geq 1-\alpha$ for all θ and for all $\theta' < \theta$, $P_\theta\{\underline{\theta}(X) \leq \theta'\} = \textit{minimum}$ is a uniformly most accurate lower confidence bound for θ at level $1-\alpha$. A family of lower confidence bounds at level $1-\alpha$ is said to be unbiased if $P_\theta\{\underline{\theta}(X) \leq \theta'\} \leq 1-\alpha$ for all $\theta' < \theta$ for all θ.

†A lower confidence bound at level $1-\alpha$ is said to be exact if $P_\theta\{\underline{\theta}(X) \leq \theta\} = 1-\alpha$ for all θ.

El Mawaziny bounds, which must be calculated iteratively from an expression which demands extremely complicated calculations when the number of subsystems is more than two or three. (Problems involving loss of precision and use of excessive amounts of computer time also arise in calculating the El Mawaziny bounds when the product of the number of independent subsystems and the number of failures for any given subsystem is more than about 50.)

Other work dealing with the derivation of confidence bounds for system reliability under the exponential-failure-time model includes a Bayesian approach for a parallel system with a single failure for each subsystem by Springer and Thompson [47] and two reports by Allen, Carlson, and Hubach [2], and Saunders [42], which discuss the fixed-test-time model for a series system.

For the case in which only pass-fail data are collected for each subsystem many methods involving large- or small-sample approximations or Bayesian techniques have been derived for obtaining confidence bounds on the probability of successful operation of an independent series system. Among the large-sample methods are those suggested by Madansky [34] (which is based on the asymptotic chi-square distribution of -2 log likelihood ratio), by Myhre and Saunders [37] (which gives a generalization of Madansky's method), and by DeCicco [11], Rosenblatt [40], and Thomas [50] (all three of which are based on the asymptotic normality of maximum-likelihood estimators). The methods of Rosenblatt and Madansky are discussed and compared by Myhre and Saunders [38], who demonstrate that the likelihood-ratio method attains its asymptotic properties for smaller sample sizes than the method suggested by Rosenblatt and in practical situations appears to yield more accurate bounds. Madansky [34], however, points out that the Rosenblatt method has slightly higher asymptotic (Bahadur) efficiency. The methods of DeCicco and Thomas use Taylor-series approximations to the variance of the maximum likelihood estimator of the system reliability, R, and would be expected to have asymptotic properties like those of the Rosenblatt method.

Small-sample approximate confidence bounds on R for an independent series system and binomial data have been derived by Nishime [39], Garner and Vail [20], Connor and Wells [8], Abraham [1] and Lindstrom and Madden (see Lloyd and Lipow [33]). The first three of these approaches use various methods of combining confidence bounds on subsystem reliability to obtain the desired bounds on system reliability. The others use binomial or Poisson approximations for certain statistics. Some of these methods are sensitive to inequality of sample sizes for subsystems. Lower confidence bounds obtained by most of these approximate methods have compared by the use of three sets of data by Schick and Prior [44] with three different sets of "exact" bounds obtained using results of Lipow (see [31] and [32]), based on Buehler's theory [6] and Poisson approximations. The data apply to systems composed of two subsystems, and in each of the three cases the sample sizes are equal. Only the Lindstrom and Madden method compares favorably with what appears to be the best of the Lipow "exact" bounds. Since there is some question about the standard used to judge the quality of the approximate methods, however, and since only three sets of data, two subsystems, and equal sample sizes have been used in the comparisons, it is very difficult to make useful general inferences concerning these results.

Another method investigated numerically by Schick and Prior [44] is the Bayesian approach wherein reliability for each subsystem is assumed to have a prior distribution which is uniform over the unit interval. Confidence bounds which are exact in the Bayesian sense (under the assumed prior distribution) are derived by Zimmer, Prairie, and Briepohl [52] and by Springer and Thompson [45]. The latter authors use a Mellin transform technique for obtaining in closed form the distribution of system reliability, given all the subsystem data. A Monte Carlo application of this Bayesian model is

suggested by Mastran [35] for a system which is logically more complex than a series system. In the numerical comparisons given by Schick and Prior [44], there appears to be no particular agreement between the sets of Bayesian bounds calculated on the basis of the procedure prescribed by Springer and Thompson and by Zimmer, et al. (which incidentally agree to two or three significant figures, as one might expect) and the three sets of "exact" bounds calculated. In particular, the Bayesian lower confidence intervals on R are all larger than those based on what is for these three sets of data the smallest of the "exact" intervals.

One would expect the Bayesian bounds to be exact in a classical sense if sample sizes for all subsystems were "sufficiently" large. This is so because a prior density of the assumed type will have less effect upon the confidence bound as the sample sizes for all subsystems increase. Whether or not the bound is exact in a classical sense has not been established. Furthermore, the accuracy of this bound has not been investigated for small sample sizes. It is interesting to note that an approximate method, described by Dalton [10] and attributed to TRW's Florida Operations, yields bounds which agree to within 3 in the third significant figure with the three examples calculated in [44] by means of this particular Bayesian approach. The TRW method has the distinction of being extremely amenable to hand calculation.

Among very recently derived approximate methods for obtaining confidence bounds on the probability of successful operation of a series system are: (1) those derived by Woods and Borsting [51] (discussed by Lieberman [30]), which are shown by Monte Carlo investigations in their paper to be very nearly exact; (2) those derived by J. R. Johnson [23] based on the exact multi-variate binomial distribution of component test data; and (3) those by J. Bram [5], arising from a Bayesian approach which formally uses subjective judgment concerning prior knowledge.

The Monte Carlo Confidence Bound Problem

We now examine the problem of obtaining lower confidence bounds on the reliability of a logically complex system when testing will be performed on the k independent subsystems only. We assume that an equation relating true subsystem reliabilities to true system reliability is available, say by means of computer programs which can provide such information (see Levy [28] and McKnight, Modiest, and Schmidt [36]). We now, in lieu of an appropriate analytical method of obtaining such bounds, consider the possibility of the use of Monte Carlo techniques as suggested by Burnett and Wales [7], Bosnikoff and Klion [4], Costello, Meisel, and Letow [9], Levy and Moore [29] and Mastran [35].

At first glance the creation of a Monte Carlo computer program for obtaining the bounds seems to be a straightforward problem of simulating the distribution of system reliability for a given set of failure data in an efficient manner. It soon becomes apparent, however, that there are important Bayesian questions implicit in the problem. That is, in order to generate the distribution of system reliability for a given data set, one must generate for each subsystem what is essentially the posterior distribution of subsystem reliability, given the subsystem life-test failure data. Hence, some prior distribution or something equivalent to such a prior distribution for subsystem reliability must be implicitly or explicitly assumed. In other words, in carrying out the Monte Carlo approach outlined by the authors mentioned above, one uses the density of some appropriate function of the data and implicitly or otherwise combines this information with a prior density of subsystem reliability by means of Bayes' Theorem,

$$P(A_i|B) = P(B|A_i)P(A_i) / \sum_{\text{all } j} P(B|A_j)P(A_j),$$

to obtain the posterior density function of subsystem reliability, given the data. In agreement with the classical analytical method derived in [49], the Monte Carlo procedures described in [4, 7, 9, 29, 35], in some cases directly suggest, and in others tacitly imply a prior distirbution for subsystem reliability which is the appropriate prior leading to the classical optimum bounds when the system consists of one subsystem only. One may then inquire as to whether such an assumption is appropriate when the system consists of more than one subsystem.

Springer and Thompson [47] analytically derive their exact Bayesian confidence bounds on $R(t_m)$ for an exponential-failure-time model, wherein one failure is allowed for each independent subsystem of a parallel system, using an alternative a priori assumption. They assume a uniform prior distribution on subsystem reliability over the unit interval, which leads to the classical optimum bounds on successful system operation for the *pass-fail* model when the system consists of a single subsystem. Springer and Thompson reason that a flat prior for subsystem reliability is in keeping with the intent of Bayes' Theorem when no prior information is known. They point out that the prior density $p(R_j)$ for the jth subsystem reliability yielding the classical optimum bounds for a system containing a single subsystem and an exponential fixed-failures model, $p(R_j) = R_j^{-1} [\ln (1/R_j)]^{-1}$ or equivalently, $q(\lambda_j) = \lambda_j^{-1}$, where $R_j \equiv R_j(t_m) = \exp(-\lambda_j t_m)$ and $0 \leq R_j \leq 1$, $j = 1, 2, \ldots, k$, is "improper" in that the area under the frequency curve cannot be made equal to unity. Mastran [35] suggests for pass-fail data that prior densities for subsystems which lead to a uniform prior density for system reliability might be appropriate. In other words all the suggested prior distributions are derived from the concept of optimality for one subsystem for *some* model, even though the model may have little relationship to the one of interest.

In the following, a description is given of results of a study (by members of a group at North American Rockwell Corporation) to determine optimum prior assumptions to be used in generating Monte Carlo confidence bounds on the reliability of a logically complex system. The investigation was conducted principally by K. W. Fertig of Rocketdyne Division and the present author. Mr. Fertig wrote all computer programs needed for the investigation, except for one routine linking the Monte Carlo program to the reliability equation for the complex system. He also provided (see [17]) the important analytical derivation of the necessary form for a special restricted model of an optimum prior density function independent of the data and proved that no such prior density exists. Jerome Spanier of the North American Rockwell Science Center provided consultation on problems related to the Monte Carlo computer program. Shirley Stoneberger of the Los Angeles Division wrote the subroutine which makes use of the reliability equation generated from engineering flow chart information by the SCOPE [28] or the ARMM [36] program for a logically complex system.

RESULTS OF INVESTIGATION

Computer Programs Written and Utilized

An optimum standard against which to judge suggested prior distributions provides a means of attacking the Monte Carlo problem. Therefore, the model was first restricted to a series system wherein the jth independent subsystem has exponentially distributed failure time, T_j, with Prob

$[T_j > t_m] \equiv R_j \equiv R_j(t_m) = \exp(-\lambda_j t_m)$ and n_j prototypes of the jth subsystem, $j=1, 2, \ldots, k$, are tested until r_j failures occur. If one can determine an appropriate prior distribution for this model, then it should also be possible to make useful inferences concerning the fixed-failure-time series-system model and to determine a method of using prior information for more complex systems.

A computer program was coded in Fortran *H* for the IBM *S*/360 system for calculating for this restricted model the optimum classical confidence bounds of Lentner, Buehler, and El Mawaziny discussed in the introduction of this paper. The bounds are based on the conditional distribution of

$$W = Z_1, \text{ given } Z_1 - Z_2 = u_2, \ldots, Z_1 - Z_k = u_k, \text{ where } Z_j = \sum_{i=1}^{r_j} T_{i,j} + (n_j - r_j) T_{r_j, j},$$

with $T_{i,j}$ an observable failure time of the ith prototype of the jth component, and where the subscript 1 is arbitrarily assigned. Then, when u_j is less than zero for $j=2, 3, \ldots, k$, the optimum classical $(1-\alpha)$-level lower confidence bound $R_B(\alpha)$ on $R(t_m) = \exp(-\phi t_m)$, (where $\phi = \sum_{j=1}^{k} \lambda_j$) is obtained by finding the solution $\phi_B(\alpha)$ of the following equation and then calculating $R_B = \exp[-\phi_B(\alpha) t_m]$, *with* $\phi_B > 0$,

$$(1)\quad H(w|u;\phi) = A^{-1}(u;\phi) \sum_{i_2} \sum_{i_3} \cdots \sum_{i_k} \prod_{j=2}^{k} \left[\binom{a_j}{i_j} (-u)^{a_j - i_j} \right] \phi^{-a_1 - \sum_j i_j - 1} \Gamma_{\phi w}(a_1 + \sum_j i_j + 1) = 1 - \alpha,$$

where

$$A(u;\phi) = \phi^{-1-\sum_1^k a_j} \sum_{i_2}^{a_2} \sum_{i_3}^{a_3} \cdots \sum_{i_k}^{a_k} (a_1 + \sum_{j=2}^{k} i_j) \prod_{j=2}^{k} \left[\binom{a_j}{i_j} (-\phi u_j)^{a_j - i_j} \right],$$

and where

$$\Gamma_{\phi w}(a_1 + \sum_j i_j + 1) = \int_0^{\phi w} y^{a_1 + \sum_{j=2}^{k} i_j} e^{-\phi y} dy.$$

A similar expression is used if any u_j, $j=2, 3, \ldots, k$, is greater than zero, and the solution is obtained by joint application of Newton-Raphson iterative procedures, the method of false position and bisection techniques.

Then a computer program for generating Monte Carlo confidence bounds was coded and combined with that for obtaining the Lentner-Buehler-El Mawaziny confidence bounds. A listing and flow chart of the combined computer program are available (see Fertig [16] and [18]).

The Monte Carlo program calculates the confidence bounds on the basis of a specified prior density for subsystem reliability which is a member of the "conjugate" family of prior densities. That is, the prior density yields a posterior density of subsystem reliability, given the subsystem data, of the same general form (belonging to the same family of density functions) as the prior density. The prior density function $p_j(R_j)$ used for the jth subsystem reliability was, therefore,

$$(2)\qquad p(R_j) = \frac{(\beta_{oj}+1)^{r_{oj}+1}}{\Gamma(r_{oj}+1)} R_j^{\beta_{oj}} [\ln(1/R_j)]^{r_{oj}}, \beta_{oj}, r_{oj} > -1, j=1, 2, \ldots, k, \text{ with } \beta_{oj} \text{ and } r_{oj},$$

subjectively chosen. This yields a posterior density $f(R_j \mid \hat{\beta}_j;\ r_j)$ for R_j of the form,

$$f(R_j \mid \hat{\beta}_j;\ r_j) = \frac{(\hat{\beta}_j + \beta_{oj} + 1)^{r_j + r_{oj} + 1}}{\Gamma(r_j + r_{oj} + 1)} R_j^{\hat{\beta}_j + \beta_{oj}} [\ln (1/R_j)]^{r_j + r_{oj}}, \tag{3}$$

where the random variable $Z_j/t_m = \left(\sum_{i=1}^{r_1} T_{i,j} + (n_j - r_j) T_{r_j,j} \right) / t_m$

is equal to $z_j/t_m = \hat{\beta}_j$, $j = 1, 2, \ldots, k$, for the observed set of data.

If β_{oj} and r_{oj} each have the value -1, then the prior density for R_j corresponds to the "improper" prior which is used by [7, 4, 29, 49], and which gives the optimal classical bounds for a system consisting of a single subsystem (see Epstein and Sobel [14.]), that is,

$$p(R_j) = R_j^{-1} [\ln (1/R_j)]^{-1}, \quad j = 1, 2, \ldots, k.$$

(It is true, therefore, that even though the prior density corresponding to $\beta_{oj} = r_{oj} = -1$ is "improper," the corresponding posterior density is proper). If β_{oj} and r_{oj} are both equal to zero, $j = 1, 2, \ldots, k$, then each subsystem prior density function for subsystem reliability is uniform over the interval from 0 to 1, as suggested by Springer and Thompson [47] for their special case of a parallel-system model mentioned earlier.

For generating the posterior distribution of R_j using the expression (3), a given set of data and specified values for β_{oj} and r_{oj}, a random number ρ_j is generated for the value of the integral $\gamma_j \equiv \gamma(R_j \mid \hat{\beta}_j;\ r_j)$ given by the expression (3) from $R_{\gamma,j}$ to 1. The integration is performed by an evaluation of the incomplete gamma function and the value of $R_{\gamma,j}$ determined iteratively. The Newton-Raphson method of iteration in conjunction with the method of false position is used. Because this procedure is quite expensive in terms of computer time, the computer program was written to calculate for a given set of subsystem data a table of 100 values of $R_{\gamma,j}$ corresponding to equally spaced values of $\gamma(R_j \mid \hat{\beta}_j;\ r_j)$. The computer then samples from and interpolates cubically in this table for $a_1 < \rho_j < a_2$, where a_1 and a_2 are functions of the data. For $\rho_j < a_1$ and $\rho_j > a_2$, a different table is sampled. In generating values for this table, $\gamma(R_j \mid \hat{\beta}_j;\ r_j)$ is calculated from a specified value of $R_{\gamma,j}$, so that no iteration is necessary, but the values of γ_j are not equally spaced (making interpolation more difficult). The second table, which contains values of γ_j much closer together than the one used for non-extreme values of $R_{\gamma,j}$, is necessary because of the steepness of the curve relating γ_j and $R_{\gamma,j}$ for values of $R_{\gamma,j}$ close to 0 or 1.

The first investigation made by means of the computer was of the two familiar prior distributions corresponding to β_{oj} and r_{oj} both equal to zero and both equal to -1, $j = 1, 2, \ldots, k$. The Bayesian approach corresponding to $\beta_{oj} = r_{oj} = -1$, incidentally, is sometimes called the fiducial model since the posterior distribution of R_j, $j = 1, 2, \ldots, k$, can be thought of as obtainable from the distribution of a function of the data for the jth subsystem, as detailed in [25]. The preliminary phases of this investigation made use of the Monte Carlo program, but the results given below were obtained using instead a computer program which utilizes a Mellin transform technique (see Springer and Thompson [46]) to calculate the posterior distribution of $R(t_m)$ from the posterior distributions of the R_j's. This Mellin transform program was originally written to calculate the variance of the Monte Carlo confidence bound and is applicable to a series system when the posterior density of R_j has the form given by the expression (3) with r_{oj} an *integer*. The Mellin transform computer program is faster than the

Monte Carlo program and gives better precision, but in its present form cannot be used if r_{oj} is other than an integer.

Study of Suggested Prior Densities

For each combination of input, involving from 3 to 25 components having λ's in various proportions, numbers of failures ranging from 1 to 10, and 3 or 4 different values of α ranging from 0.05 to 0.50, data were generated and a comparison was made of the two Bayesian bounds with the optimum classical bound. In *each* case (of a total of 156 cases), the Bayesian bound based on $\beta_{oj}=r_{oj}=-1$ is smaller than the corresponding classical bound obtained. It therefore appears that though exact in a Bayesian sense (under the assumed prior distribution for R_j, $j=1, 2, \ldots k$), the bounds based on such a prior assumption are conservative in the classical sense.

When the optimum bound is standardized at 0.800 by adjusting the mission time, t_m, the fiducial bound ranges from 0.538 to 0.793. When the optimum bound is equal to 0.368, the fiducial bound ranges from 0.062 to 0.354.

El Mawaziny and Buehler [13] show that their large-sample approximation of the optimal bound, a bound otained by the Roseblatt method [40] and the fiducial bound will approach the optimal bound as numbers of failures for all subsystems become large. For three samples having 10 failures for each of 3 identical components, the fiducial bounds were of the order 0.787 and 0.341 for optimal bounds of 0.800 and 0.378, respectively, with deviation between any two corresponding fiducial bounds less than 3 in the third decimal place.

Analytical results described later indicate that the fiducial bounds for a fixed number of failures per subsystem will agree less well as the number of subsystems increases and the subsystems become more variable with respect to failure rate. Unfortunately, because of the computer-time factor and considerations of precision, it is impossible at present to compare bounds for systems containing as many as 10 subsystems when as many as 10 failures occur for more than 1 or 2 of these subsystems. In any case, the large-sample methods cannot be expected to give bounds agreeing well with the optimal bounds when some of the subsystems have been subjected to few tests. Furthermore, it is impossible on the basis of these results to say whether bounds based on this specified prior might be conservative, liberal, or exact for a particular logically complex system.

The uniform prior distribution for subsystem reliability gives bounds even lower than those based of the fiducial method except in 24 cases (out of 150) in which all three bounds have values fairly close to zero. In these 24 cases they are higher than the optimum classical bounds. It appears that the distribution of the bounds based on the uniform prior may be less disperse than those of the optimal bounds, but these bounds seem to be even more conservative than the "fiducial" bounds given by $\beta_{oj}=r_{oj}=-1$ for true reliabilities of a reasonable size and α's of interest. For systems with low reliabilities, bounds obtained using a uniform prior density for subsystem reliability should be liberal rather the conservative when the confidence level is sufficiently low, but not exact in general. This inconsistent behavior may be due to the fact that a uniform prior density for R_j implies a prior density for λ_j (the failure rate for the jth subsystem) of the form $q(\lambda_j)=t_m \exp(-\lambda_j t_m)$, $j=1, 2, \ldots, k$, or, strangely, one which is a function of t_m, the specified mission time.

The result for $\beta_{oj}=r_{oj}=-1$ is keeping with the analysis and numerical results of Saunders [42], who studies a fixed-test-time exponential series-system model and the Bayesian approach suggested in [2]. This Bayesian approach uses a prior density for the fixed-test-time model equivalent to the so-called fiducial method. Saunders [42] points out that in using such a Bayesian model for an exponential series system (and his argument applies to any true Bayesian model, that is, one based on a prior as-

sumption which does not involve information concerning the number of components in the system), one can obtain different confidence bounds depending upon what one chooses to call a subsystem. Saunders points out, too, that such inferences apply to more logically complex systems which are highly reliable, since one can approximate an extremely reliable coherent system (see Birnbaum, Esary, and Saunders [3] for a definition of a coherent system) by a series-system model, as indicated by Esary, Proschan, and Walkup [15].

The Search for Optimum Prior Assumptions

Initially it has been planned that a trial and error procedure would be used in attempting to determine appropriate prior assumptions for our series-system model. Saunders' argument might lead one to consider trying prior assumptions which are not truly Bayesian in that they are dependent upon the configuration (or number of components in the system). At this point in the study, however, an analytical result was derived, modifying the subsequent approach. The details of the analysis are given by Fertig [17] and are summarized below.

First, the form of a prior density function, or generalization of such a function, for $R_j, j=1, 2, \ldots, k$, corresponding to the optimum classical bounds for our system model was determined. This was accomplished by setting the Laplace transform of $H_\phi(w|u;\phi)$ equal to the Laplace transform of the posterior density of

$$\phi=\sum_{j=1}^{k}\lambda_j,$$

obtained under a general prior assumption (*not restricted to conjugate priors*) for the special case $u_2=u_3=\ldots=u_k=0$. If the prior assumptions which yield the optimum bounds are independent of the data, true Bayesian priors, for example, then an assumption concerning the value of the u's will have no effect on the result. The fact that the optimum classical bounds are invariant under permutations of the subsystems was used to obtain the improper "prior density" for the jth subsystem yielding these bounds. It is

$$p(R_j)=R_j^{-j}\,[\ln\,(1/R_j)]^{-(2-1/k)},\ j=1, 2, \ldots, k.$$

We note that this improper density depends upon k, the number of subsystems in the series system and for $k=1$ *does* yield the optimum classical bound.

The Monte Carlo program was then used to test whether this prior assumption (which we may think of as a weighting function since it does not correspond to a strict Bayesian prior density) would yield the optimum bounds for variations in the data. For the case where all the u's equal zero ($z_1=z_2=\ldots=z_k$), 8 values of the Monte Carlo bounds based on 5,000 replications agree with the optimum bounds to within 3 in the third decimal place. Data were randomly generated for five subsystems using the fact that $2Z_j/\lambda_j$ is distributed as chi-square with $2r_j$ degrees of freedom (see Epstein and Sobel [14]), where $r_1=3$, $r_2=4$, $r_3=4$, $r_4=2$, $r_5=2$, $\lambda_1=1/12$, $\lambda_2=1/13$, $\lambda_3=1/15$, $\lambda_4=1/10$, $\lambda_4=1/11$. The mission time was taken as 1.0. For 10 such data sets the Monte Carlo bounds are uniformly larger than the classical confidence bounds with deviations ranging from 1 to 6 in the second decimal place.

The confidence bound obtained from any set of data by El Mawaziny's formula given by Eq. (1) is the *unique* optimum (uniformly most accurate unbiased) confidence bound for this exponential-failure-

number series-system model. This is proved in the Appendix of Fertig's paper. Thus, any optimum bound is equivalent to the bound defined by Eq. (1) and must give the same result for any given set of data. Since the error in the Monte Carlo procedure is very small compared with the deviations obtained for the u's not all equal to zero, the empirical evidence indicates that the prior assumptions which yield the optimum classical bounds *do* depend upon the data.

Fortunately, a means of proving this result analytically then presented itself (see Fertig [17] for details). The Laplace transform of $H(w|u;\ \phi)$ is an horrendous expression which gives no apparent clue as to how it might be factored and assigned to the various subsystems. The problem was made tractable earlier by letting all the z's be equal.

Another method of simplifying the expression was found to be to assume $r_1=r_2=\ .\ .\ .\ =r_k=1$. If this is done, then it is possible to demonstrate that the "prior density" for the jth subsystem yielding the optimum bound for $r_1=r_1=\ .\ .\ .\ =r_k=1$ cannot have the form $p(R_j)=R_j^{-1}[\ln\ (1/R)\ ]^{-(2-1/k)}$ unless $z_1=z_2=\ .\ .\ .\ =z_k$, as assumed in the earlier case. Hence, as indicated by the numerical evidence, one must incorporate present data into the "prior assumptions" or more properly the weighting functions, for obtaining optimum confidence bounds for the exponential fixed-failure-number series-system model.

Now that we have established what is not fruitful for obtaining the optimal bounds, one may properly inquire as to the next step in the investigation with respect to confidence for a complex system. Two approaches present themselves. The first is to consider each series system which is a part of the complex system as a single subsystem of the total system. If the fiducial approach were to be used, one could obtain an estimate $R_{\rho,\,l}$ of true reliability, given the failure data, for the lth subsystem consisting of the k_l independent subsystems making up this series-system subsystem by substituting a random number ρ_l for $1-\alpha$ in Eq. (1). Then the reliability equation could be used to obtain an estimate of total system reliability given the failure data. This method will probably not yield confidence bounds that are exact in the classical sense, though of course they are exact in a Bayesian sense. In lieu of the fiducial approach one could somehow modify the value obtained for $R_{\rho,\,l}$ in an attempt to obtain bounds exact in the classical sense. Clues as to how this might be accomplished may possibly be obtained by investigating a simple parallel system (again an exponential-failure-time model with fixed number of failures) and methods of obtaining confidence bounds exact in the classical sense for the simpler model.

A considerable amount of computer time will be required with this approach when the product of the number of subsystems in any series system and the number of failures for any subsystem in that series system becomes large. Hence, one might in such cases, abandon the idea of considering the series system as a single subsystem. Instead one might consider a method of approximating the optimum confidence bound for a series system by using series system data in the prior assumptions (or weighting functions) for the subsystems of the series system that yield the posterior distribution of series system reliability. For the case $r_1=r_2=\ .\ .\ .\ r_k=1$, one possible factoring of the Laplace transform gives

$$p_l(R_j)=R_j^{1-\hat{\beta}_j+\hat{\beta}_{(1)}}\,[\ln\ (1/R_j)\ ]^{-(2-1/k_l)},\qquad j=1,\ 2,\ .\ .\ .\ k_l, \tag{4}$$

where $\hat{\beta}_{(1)}$ is the smallest of $z_1/t_m,\ z_2/t_m,\ .\ .\ .\ z_{k_l}/t_m$. This function or some modification of this function involving the true values for $r_1,\ r_2,\ .\ .\ .,\ r_{k_l}$ could be tried. If the expression (4) is used without modification, then the posterior distribution of series system reliability, given the failure data, depends only upon $\hat{\beta}_{(1)}$. It can be shown that, in fact, the fiducial distribution of series system reliability for this model depends only upon $\hat{\beta}_{(1)}$ if and only if $r_1=r_2=\ .\ .\ .=r_{k_l}=1$.

The method of Kraemer [24] depends solely upon $\hat{\beta}_{(1)}$ and hence would be expected to give poor results for large numbers of failures per component. This is borne out by the comparisons made by Sarkar [41] and Grubbs [21].

The bounds derived by Grubbs [21] are approximations to the fiducial bounds, and for the bounds compared during this study, the approximation appears to be excellent. The Grubbs method is based on the fact that $2r_j\lambda_j/\hat{\lambda}_j$ is distributed as chi-square with $2r_j$ degrees of freedom, $j=1, 2, \ldots, k$, so that

$$\phi = \sum_{i=1}^{k} \lambda_j$$

can be thought of as distributed as a weighted sum of chi-square variates. The weights used by Grubbs, namely $\hat{\lambda}_j/2r_j$, $j=1, 2, \ldots, k$, are appropriate for obtaining the fiducial bounds. One could obtain, instead, an approximation to the optimal bounds by adjusting the weights appropriately, obtaining clues from the expression (4) above. In this way, one can test approximations to the optimum prior assumptions and avoid the time-consuming Monte Carlo calculations. If successful in closely approximating the optimum bounds by the proper modification of the Grubbs method, one can use this approximation in place of the Lentner-Buehler-El Mawaziny bounds in considering series systems within a complex system as single subsystems.

The investigation is being continued along these lines.

REFERENCES

[1] Abraham, John K., "A Confidence Interval for the Reliability of Multi-Component Systems," Proceedings Seventh Conference on the Design of Experiments in Army Research, Development and Testing, ARODR 62–2 (1962), pp. 483–518.

[2] Allen, D. C., C. H. Carlson, and C. E. Hubach, "Procedure for Reliability Assessment and Confidence," Federal Electric Corporation Document No. AS–A–86–67 (1967).

[3] Birnbaum, Z. W., J. D. Esary, and S. C. Saunders, "Multi-Component Systems and Structures and Their Reliability," Technometrics *3*, 55–71 (1961).

[4] Bosnikoff, I. and J. Klion, "Development of New Prediction Techniques," Proceedings Eighth National Symposium on Reliability and Quality Control (1962), pp. 382–387.

[5] Bram, J., "Confidence Limits for System Reliability," OEG Research Contribution No. 79, Center for Naval Analyses, University of Rochester (1969).

[6] Buehler, R. J., "Confidence Intervals for the Product of Two Binomial Parameters," J. Am. Statist. Assoc. *52*, 482–493 (1957).

[7] Burnett, Thomas L. and Beverly A. Wales, "System Reliability Confidence Limits," Proccedings Seventh National Symposium on Reliability and Quality Control (1961), pp. 118–128.

[8] Conner, W. S. and W. T. Wells, "Simulating Tests of a System from Tests of Its Components," Proceedings of the Eighth National Symposium on Reliability and Quality Control (1962), pp. 14–16.

[9] Costello, D. L., R. M. Meisel, and A. M. Letow, "Compliance Demonstration of a Multimode System Using Monte Carlo Analysis," Proceedings Eighth National Symposium on Reliability and Quality Control (1962), pp. 446–457.

[10] Dalton, R. E., "An Evaluation of Methods for Construction of Confidence Limits for System Reliability," TRW Systems, Florida Operations, Contract AF04 (694)–806 (1966).

[11] DeCicco, Henry, "The Error in Linearized Estimates of the Variance of Products," Technical

Note No. 2, Office for Reliability, Research and Effects (Reliability Branch—ORDSW—DR) Ordnance Special Weapons Ammunition Command Picatinny Arsenal, Dover, N.J. (1960).

[12] El Mawaziny, A. H., "Chi-Square Distribution Theory with Applications to Reliability Problems," Ph. D. Thesis, Iowa State University, Ames (1965).

[13] El Mawaziny, A. H. and R. J. Buehler, "Confidence Limits for the Reliability of a Series System," J. Am. Statist. Assoc. *62*, 1452–59 (1967).

[14] Epstein, B. and M. Sobel, "Life Testing," J. Am. Statist. Assoc. *48*, 486–502 (1963).

[15] Esary, J. D., Frank Proschan, and D. W. Walkup, "A Multivariate Notion of Association, with a Reliability Application," Boeing Document D1–82–0567 (1966).

[16] Fertig, K. W , "Flow Chart for Monte Carlo Program for Confidence Bounds on System Reliability," Rocketdyne Report MSM 69–11 (1969).

[17] Fertig, K. W., "A Result Concerning Bayesian Prior Distributions and Confidence Bounds on the Reliability of Serial Systems with Exponential Failure Times," Rocketdyne Research Report RR 69–6 (1969). (Submitted for publication.)

[18] Fertig, K. W., "Monte Carlo Program for Confidence Bounds on System Reliability," Rocketdyne Report MSM 69–12 (1969).

[19] Garner, Norman R., "Estimation for Serially Connected Systems," Technical Report RCS59–4 Aerojet-General Corporation (1969).

[20] Garner, Norman R. and Richard W. J. Vail, "Confidence Limits for System Reliability," Military Systems Design Vol. 7, No. 5 (1961).

[21] Grubbs, Frank, "On Confidence Limits for the Reliability of a Series System for Which Each Component has an Exponential Time-to-Fail Distribution," (1969). (Submitted for publication.)

[22] Harris, Bernard, "Hypothesis Testing and Confidence Intervals for Products and Quotients of Poisson Parameters with Applications to Reliability," U.S. Army Mathematics Research Center, The University of Wisconsin, Madison, Wisconsin, MRC Technical Summary Report No. 923 (1968).

[23] Johnson, J. R., "Confidence Interval Estimation of the Reliability of Multicomponent Systems Using Conponent Test Data," Ph.D. Thesis, University of Delaware (1969).

[24] Kraemer, H. C., "One-Sided Confidence Intervals for the Quality Indices of a Complex Item," Technometrics *5*, 400–403 (1963).

[25] Kendall, M. G. and A. Stuart, *The Advanced Theory of Statistics* (Hafner Pub. Co., N.Y., 1967), Vol. 2, 2d ed.

[26] Lehman, E. L., *Testing Statistical Hypotheses* (John Wiley, N.Y., 1959).

[27] Lentner, M. M. and R. J. Buehler, "Some Inferences about Gamma Parameters with an Application to a Reliability Problem," J. Am. Statist. Assoc. *58*, 670–677 (1963).

[28] Levy, Sherwin, "System to Compute Operational Probability Equation," Program XC0003, Space Division, North American Rockwell Corporation (1969).

[29] Levy, Louis L. and Albert H. Moore, "A Monte Carlo Technique for Obtaining System Reliability Confidence Limits from Component Test Data," I.E.E.E. Transactions on Reliability *R–16*, 69–72 (1967).

[30] Lieberman, Gerald J., "The Status and Impact of Reliability Methodology," *Naval Research Logistics Quarterly 16*, 17–35 (1969).

[31] Lipow, M., "Measurement of Over-All Reliability Utilizing Results of Independent Subsystem Tests," GM–TR–0165–00506, Space Technology Laboratories (1958).

[32] Lipow, M., "Tables of Upper Confidence Bounds on Failure Probability of 1, 2, and 3 Component Serial Systems," TR-50-0000-00756, Space Technology Laboratories (1959) (2 Vols.).

[33] Lloyd, D. K. and M. Lipow, *Reliability: Management, Methods, and Mathematics* (Prentice-Hall, N.J., 1962).

[34] Madansky, Albert, "Approximate Confidence Limits for the Reliability of Series and Parallel Systems," Technometrics 7, 495-503 (1965).

[35] Mastran, David V., "A Bayesian Approach for Assessing the Reliability of Air Force Re-entry Systems," Proceedings of the ASME Reliability and Maintainability Symposium (1968), pp. 380-383.

[36] McKnight, C. W., L. J. Modiest, and N. E. Schmidt, "An Automatic Reliability Mathematical Model," Proceedings Eleventh National Symposium on Reliability and Quality Control (1965), pp. 518-532.

[37] Myrhe, Janet and Sam C. Saunders, "On Confidence Limits for the Reliability of Systems," Ann. Math. Statist. *39*, 1463-1472 (1965).

[38] Myhre, J. M. and Sam C. Saunders, "Comparison of Two Methods of Obtaining Approximate Confidence Intervals for System Reliability," Technometrics *10*, 37-49 (1968).

[39] Nishime, F., "Techniques for . . . the Establishment of Confidence Limits for the Estimated Reliability" (unpublished memorandum), Space Technology Laboratories (1959).

[40] Rosenblatt, Joan Raup, "Confidence Limits for the Reliability of Complex Systems," *Statistical Theory of Reliability* (Ed. Marvin Zelen) (The University of Wisconsin Press, Madison, 1963), pp. 115-137.

[41] Sarkar, Tapas K., "An Exact Lower Confidence Bound for the Reliability of a Series System Where Each Component Has an Exponential Time to Failure Distribution," Department of Operations Research and Department of Statistics, Stanford University, Technical Report No. 117 (1969).

[42] Saunders, Sam C., "On Confidence Limits for the Performance of a System When Few Failures are Encountered," Boeing Document D1-82-0676 (1969), Revised.

[43] Schick, G. J., "Reliabilities, Confidence Limits and Their Improvements as Applied to Missile Reliability," Technical Publication, Aerojet General Corporation (1959).

[44] Schick, G. J. and R. J. Prior, "Reliability and Confidence of Serially Connected Systems," Proceedings of the Third Space Congress, Cocoa Beach, Florida (1966), pp. 352-360.

[45] Springer, M. D. and W. E. Thompson, "Bayesian Confidence Limits for the Product of n Binominal Parameters," Biometrika *53*, 611-613 (1966).

[46] Springer, M. D. and W. E. Thompson, "Bayesian Confidence Limits for System Reliability," Proceedings of the ASME Reliability and Maintainability Symposium (1969), pp. 515-523.

[47] Springer, M. D. and W. E. Thompson, "Bayesian Confidence Limits for Reliability of Redundant Systems When Tests Are Terminated at First Failure," Technometrics *10*, 29-36 (1968).

[48] Steck, G. P., "Upper Confidence Limits of the Failure Probability of Complex Networks," Sandia Corporation, SC-433 (TR) (1957).

[49] Takenaga, R., "Predicting System Reliability with Associated Confidence Level from Component Test Data," Technical Memorandum 3024-43-MA-003, Autonetics, A Division of North American Aviation, Inc., Downey, California (1962).

[50] Thomas, Ralph E., "An Improved Formula for the Standard Deviation of the Reliability Product Rule," Battelle Memorial Institute, Technical Report No. 2 (1960).

[51] Woods, W. Max and J. R. Borsting, "A Method for Computing Lower Confidence Limits on System

Reliability Using Component Failure Data with Unequal Sample Sizes," United States Naval Postgraduate School, Monterey, California, NPS 55Wo/Gg 8061A (1968).

[52] Zimmer, W. J., R. R. Prairie, and A. M. Breipohl, "A Consideration of the Bayesian Approach in Reliability Evaluation," I.E.E.E. Transactions on Reliability *R–14*, 107–113 (1965).

* * *

A SEQUENTIAL BAYES PROCEDURE FOR RELIABILITY DEMONSTRATION

Ray E. Schafer and Nozer D. Singpurwalla*

Hughes Aircraft Company, Systems Effectiveness Department, Fullerton, California

ABSTRACT

A common problem in life testing is to demonstrate that the mean time to failure, θ, exceeds some minimum acceptable value, say θ_1, with a given confidence coefficient γ. When this is true, it is said that "θ_1 has been demonstrated with a confidence γ." In this paper a Sequential Bayes Procedure (SBP) for demonstrating (by means of a probability statement) that θ exceeds θ_1 is presented. The SBP differs from the classical procedure in the sense that a prior distribution is assumed on the parameter θ, calling for a Bayesian approach. The procedure is based on the sequence of statistics

$$\Pr\ (\theta \geq \theta_1 | \hat{\theta}_n),$$

where $\hat{\theta}_n$ is an estimator of θ obtained at the nth step of the sequential procedure. The paper also develops bounds on the quantiles of the distribution of the number of steps to termination, and an approximation to the probability of acceptance. The exponential failure law is considered with an inverted gamma prior and a two point prior distribution, respectively, on the scale parameter.

1. INTRODUCTION

Let θ be a single parameter in a continuous probability density function, $f(t; \theta)$, for a random variable t. The classical procedure for testing the hypotheses

$$H_0 : \theta = \theta_0 \text{ versus}$$

$$H_1 : \theta = \theta_1 < \theta_0$$

is to find a fixed sample size test (or a sequential test), such that

$$Pr(\text{Reject } H_0 | \theta = \theta_0) = \alpha, \text{ and}$$

$$Pr(\text{Accept } H_0 | \theta = \theta_1) = \beta,$$

where α and β are small numbers. The 4-tuple (θ_0, θ_1, α, β) uniquely determines the sample size n and also leads to a criterion for deciding whether to accept or to reject H_0 based on sample data. The reliability analogue of the above test of a hypothesis arises when t is time to failure, and θ, θ_0, and θ_1 are the values of the mean time between failure (MTBF). α and β are commonly referred to as the producer's and the consumer's risks, respectively. It is customarily stated in reliability work, when such a test is passed, i.e., when H_0 is accepted, then θ_1 has been demonstrated with $\gamma = 1 - \beta$ confidence. This is

*Presently on a leave of absence with the Institute for Management Science and Engineering, The George Washington University, Washington, D.C.

because of the fact that if θ were less than θ_1, the chances of passing the test would be small ($< \beta$), and hence θ must have been larger than θ_1. When θ_1 has been demonstrated to a $1-\beta$ confidence, it is often implied that

$$Pr(\theta \geq \theta_1) \geq 1-\beta,$$

even though this is not necessarily true. This is because in the classical procedure θ is not a random variable and therefore one cannot make a probability statement about it. This is one of the reasons why a Bayesian approach has been advocated instead of a non-Bayesian approach. θ_1 is called the lower confidence limit and $1-\beta$ is called the confidence coefficient.

The present paper addresses itself to a method which provides a direct probability statement. The consumer specifies a minimum acceptable MTBF (say θ_1), and would like a test procedure which is based on all available evidence, and which, on acceptance, guarantees that the probability that θ exceeds θ_1 is high. The probability that $(\theta \geq \theta_1)$ is conditioned by the sample evidence (the sample statistic $\hat{\theta}_n$) and is referred to as a posterior probability. In order to compute this posterior probability, a prior distribution on θ is needed. Alternately, the producer wants a test procedure, which, on rejection, guarantees that the posterior probability that θ exceeds θ_1 is low. The above two criteria are a direct consequence of the concept of Posterior Consumer's Risk and Posterior Producer's Risk, respectively, and for a detailed discussion of these the reader is referred to [5].

Note that the classical procedure does not include the use of a prior distribution on θ, even though in many cases the existence of a prior distribution on θ can be justified [3].

This paper presents a sequential procedure which "demonstrates" θ_1 to a high probability in a sense that when acceptance has occurred, $Pr(\theta \geq \theta_1|\hat{\theta}_n)$ is very high, and when rejection has occurred, $Pr(\theta \geq \theta_1|\hat{\theta}_n)$ is very low.

2. THE SEQUENTIAL BAYES PROCEDURE

Let $f(t|\theta)$ be the probability density function (p.d.f.) of the time to failure random variable t of a unit of a product possessing a fixed but unknown mean time between failure (MTBF) θ. Let $g(\theta)$ be a p.d.f. on θ which shall be referred to as a prior density on θ.

It is assumed that a value θ is drawn according to $g(\theta)$ as the test is begun, and that each observed lifetime, say t, is a random sample from $f(t|\theta)$. The basic problem here is that a θ has been selected from $g(\theta)$, and one is attempting to decide whether the θ belongs to the set $(\theta \geq \theta_1)$ or to the set $(\theta < \theta_1)$. θ_1 can be looked upon as representing the minimal acceptable MTBF, or a lower confidence limit.

The Sequential Bayes Procedure (SBP) allows one to conduct testing, until either one of the following two criteria are satisfied:

$$Pr(\theta \geq \theta_1|\hat{\theta}_n) \geq K_2,$$

in which case a decision to accept H_0 is made, or

$$Pr(\theta \geq \theta_1|\hat{\theta}_n) \leq K_1 \qquad (0 < K_1 < K_2 < 1)$$

in which case a decision to reject H_0 is made.

The choice of K_2 and K_1 will have to be negotiated between the consumer and the producer. K_2 can be identified with the consumer's risk $(1-\beta)$ and K_1 with the producer's risk α. A discussion on the choices of K_2 and K_1 is deferred to a later section (see Section 2.2), but presently it may suffice to state that a large value of K_2 will be desired by the consumer, say about 0.9 or so, and a small value of K_1, say about 0.05 or so, is desired by the producer.

Let $\hat{\theta}_n$ be the sample MTBF based on n observed lifetimes t_i. In general (see Equation 3.1) the p.d.f. of $\hat{\theta}_n$ depends on θ. Let $f_n(\hat{\theta}_n|\theta)$ be the p.d.f. of $\hat{\theta}_n$ conditioned on θ. Using the theorem on conditional probabilities, it easy to see that the p.d.f. of θ conditioned on $\hat{\theta}_n$ is

$$g_n(\theta|\hat{\theta}_n)=\frac{f_n(\hat{\theta}_n|\theta)g(\theta)}{f_n(\hat{\theta}_n)} \qquad (f_n(\hat{\theta}_n)>0), \tag{2-1}$$

where $f_n(\hat{\theta}_n)$ is the unconditional p.d.f. of $\hat{\theta}_n$ and is given by

$$f_n(\hat{\theta}_n)=\int_0^\infty f_n(\hat{\theta}_n|\theta)g(\theta)d\theta.$$

Now consider the sequence of statistics $P_n=Pr(\theta\geq\theta_1|\hat{\theta}_n)$. From Equation (2-1) it is clear that

$$P_n=\int_{\theta_1}^\infty\left[\frac{f_n(\hat{\theta}_n|\theta)g(\theta)}{f_n(\hat{\theta}_n)}\right]d\theta.$$

The (SBP) is based on the sequence of statistics P_n. Having selected K_2 and K_1, the procedure is applied as follows:

*STEP 1** – Compute

$$Pr(\theta\geq\theta_1)=\int_{\theta_1}^\infty g(\theta)d\theta=P_0.$$

If $P_0\geq K_2$, accept the product and do not test, and if $P_0\leq K_1$, reject the product and do not test.

STEP 2 – If $K_1<P_0<K_2$, observe a lifetime t, and compute

$$P_1=Pr(\theta\geq\theta_1|\hat{\theta}_1),$$

if $P_1\geq K_2$, terminate the test by acceptance,
if $P_1\geq K_1$, terminate the test by rejection, and,
if $K_1<P_1<K_2$, proceed to the next step.

STEP n – At the nth stage compute

$$P_n=Pr(\theta\geq\theta_1|\hat{\theta}_n),$$

where $\hat{\theta}_n$ is the estimate of θ computed from n failure times $(n=2, 3, \ldots)$.

If $P_n\geq K_2$, terminate the test by acceptance,
if $P_n\leq K_1$, terminate the test by rejection, and,
if $K_1<P_n<K_2$, observe another lifetime and compute P_{n+1}.

---

*Step 1 will be referred to as the case of "no testing." It is evident that if the prior distribution is such that if $Pr(\theta\geq\theta_1)$ is near K_2 or K_1, the test time will, in general, be short.

Some interesting observations on the SBP are in order:

i) If a decision to accept has been made at the nth stage, then $P_n = Pr(\theta \geq \theta_1 | \hat{\theta}_n) \geq K_2$ as promised; if at the nth stage a decision to reject has been made, then $P_n = Pr(\theta \geq \theta_1 | \hat{\theta}_n) \leq K_1$ as promised.

In other words, when a termination has resulted due to acceptance, there is a high($\geq K_2$) probability that indeed $\theta \geq \theta_1$, and when a rejection has occurred, there is a low probability ($\leq K_1$) that $\theta \geq \theta_1$.

ii) It is to be noted that the computation of P_n at each stage is not necessary. This is because, prior to the commencement of the test, two sets S_{ln} and S_{un} can be computed for each n, such that $\hat{\theta}_n \epsilon S_{un}$ implies that $P_n \geq K_2$ and conversely, and $\hat{\theta}_n \epsilon S_{ln}$ implies that $P_n \leq K_1$ and conversely. Thus, at the nth step, $\hat{\theta}_n$ need only be tested for membership in S_{ln} or in S_{lu}. If $\hat{\theta}_n$ belongs to the former the test is discontinued resulting in a rejection and if $\hat{\theta}_n$ belongs to the latter the test is discontinued resulting in an acceptance. If $\hat{\theta}_n$ belongs to neither, testing is continued.

If P_n is monotonic in $\hat{\theta}_n$, then the sets S_{ln} and S_{lu} will be of the form $(0, \hat{\theta}_{n*})$ and $(\hat{\theta}_n^*, \infty)$, respectively.

2.1 Termination of the SBP

For any sequential procedure to be meaningful it is necessary to prove that the procedure does not continue indefinitely but does in fact terminate. After presenting a set of sufficient conditions under which P_n, converges in probability to 1 or to 0, it has been shown by Bricker and Schafer [1], that the SBP terminates with a probability of 1 either via an acceptance of H_0, or via its rejection under very general conditions. In the interest of brevity the proof has been omitted here.

The sufficient conditions are that

i) the mean of the posterior distribution of θ converges in probability to the unknown selected value of θ, and that

ii) the variance of the posterior distribution of θ converges in probability to zero.

2.2 An Approximation for the Probability of Acceptance

The probability of acceptance $P(A)$ is given by an infinite summation of an expression of the type given below.

$$P(A) = Pr(P_0 \geq K_2) + Pr(K_1 < P_0 < K_2,\ P_1 \geq K_2)$$

$$+ Pr(K_1 < P_0 < K_2,\ K_1 < P_1 < K_2,\ P_2 \geq K_2)$$

$$+ Pr(K_1 < P_0 < K_2,\ K_1 < P_1 < K_2,\ K_1 < P_2 < K_2,\ P_3 \geq K_3)$$

$$+ \ldots$$

where

$$P_0 = Pr(\theta \geq \theta_1), \text{ and}$$

$$P_i = Pr(\theta \geq \theta_1 | \hat{\theta}_i).$$

This form is certainly intractable even though $Pr(P_n \geq K_2)$ and $Pr(K_1 < P_n < K_2)$ can be computed from $f_n(\hat{\theta}_n)$. To circumvent this problem, the following approximation for $P(A)$ has been developed.

Since P^{ν} is a random function of $\hat{\theta}_n$, the expectation of P_n is given by

$$E(P_n) = \int_0^\infty P_n f_n(\hat{\theta}_n) d\hat{\theta}_n$$

$$= \int_0^\infty \left[\int_{\theta_1}^\infty \frac{f_n(\hat{\theta}_n | \theta) g(\theta)}{f_n(\hat{\theta}_n)} d\theta \right] f_n(\hat{\theta}_n) d\hat{\theta}_n$$

$$= \int_0^\infty \int_{\theta_1}^\infty [f_n(\hat{\theta}_n | \theta) g(\theta)] d\theta d\hat{\theta}_n$$

$$= \int_{\theta_1}^\infty \int_0^\infty [f_n(\hat{\theta}_n | \theta) g(\theta)] d\hat{\theta}_n d\theta.$$

Since $\int_0^\infty f_n(\hat{\theta}_n | \theta) d\hat{\theta}_n = 1$,

$$E(P_n) = \int_{\theta_1}^\infty g(\theta) d\theta = P(\theta \geq \theta_1) = P_0. \tag{2-2}$$

It is to be noted that $E(P_n)$ is independent of n. Now suppose that when the test is terminated, the excess of P_n over K_2 or the deficit of P_n under K_1 is neglected. That is, it is assumed that at termination either $P_n = K_2$ or $P_n = K_1$. Then

$$E(P_n) = P(\theta \geq \theta_1) \cong K_2 P(A) + K_1[1 - P(A)].$$

Therefore,

$$P(A) \equiv 0, \qquad 0 \leq P(\theta \geq \theta_1) \leq K_1.$$

$$P(A) \cong \frac{P(\theta \geq \theta_1) - K_1}{K_2 - K_1}, \qquad K_1 < P(\theta \geq \theta_1) < K_2, \tag{2-3}$$

$$P(A) \equiv 1, \qquad K_2 \leq P(\theta \geq \theta_1) \leq 1.$$

The quality of this approximation has not been analytically justified. However, an empirical investigation of the goodness of the approximation to $P(A)$ was undertaken (see Section 4), and this indicates that the approximation yields satisfactory results. Since the above approximation involves assuming a two point density on P_n ($P_n = K_2$ or K_1), instead of a continuous density, the approximation will be better for tests which need a large number of steps for termination. Thus when $P(\theta \geq \theta_1)$ is near K_2 or near K_1 (short tests), the approximation will not be as good when $P(\theta \geq \theta_1)$ is near the mid-point of K_2 and K_1. This occurs because when $P(\theta \geq \theta_1)$ is at the mid-point of K_2 and K_1, the approximation tends to smooth out the excess of P_n over K_2 and the deficit of P_n under K_1. The approximation is at its best when $P(A)$ is not near one or zero (i.e., P_n near K_2 or near K_1). This should not be of concern, since it is those values of $P(A)$ which are moderate that are of prime interest. The empirical investigation did confirm the above facts.

Equation (2–3) brings out another point. In Section 2 the choice of K_2 and K_1 was mentioned. It is likely in practice, that the consumer will select a K_2, and the producer instead of selecting a K_1, will select a desired value of $P(A)$ and will solve for K_1 using Equation (2–3).

2.3 Bound on the Percentiles of the Distribution of the Number of Observations to Termination

Let n_t be the random variable denoting the number of observations required to reach a decision.

Then, the expectation of n_t is given by

$$E(n_t) = \sum_{i=1}^{\infty} i\{Pr(P_i \geq K_2, K_1 < P_{i-1} < K_2, \ldots, K_1 < P_1 < K_2) + Pr(P_i \leq K_1, K_1 < P_{i-1} < K_2, \ldots, K_1 < P_1 < K_2)\}.$$

Again, this form is intractable even though the individual probabilities can be computed from $f_n(\hat{\theta}_n)$. It turns out, however, that a lower bound on the percentiles of the distribution function of n_t is available with much less computational effort.

The bound is given by the expression

$$Pr(n_t \leq n) \geq Pr(P_n \geq K_2) + Pr(P_n \leq K_1).$$

The proof of the above expression is as follows. Let n be arbitrary (i.e., n may be even larger than n_t). Now, the event

$$(P_n \geq K_2) \cup (P_n \leq K_1) \text{ implies the event } n_t \leq n.$$

$$\therefore Pr(n_t \leq n) \geq Pr(P_n \geq K_2) + Pr(P_n \leq K_1), \text{ since}$$

the events $(P_n \geq K_2)$ and $(P_n \leq K_1)$ are mutually exclusive.

$Pr(P_n \geq K_2)$ and $Pr(P_n \leq K_1)$ can be easily computed from $f_n(\hat{\theta}_n)$.

For example if for an $n = 20$, it is found that $Pr(P_n \geq K_2) + Pr(P_n \leq K_1) = 0.90$, then this implies that the test would terminate on or before 20 observations with a probability of at least 0.90.

3. AN APPLICATION OF THE RESULTS TO SPECIFIC CASES

In this section, the results developed in Section 2 will be applied to specific conditional and prior densities. The exponential conditional density is chosen because of its simplicity and usefulness. For illustrative purposes, an inverted gamma density and a two-point density are chosen as prior densities on the scale parameter of the exponential conditional density. This by no means is a limitation, since the results are general enough for extension to other single parameter densities.

3.1 The Exponential Conditional Distribution

It is assumed that

$$f(t|\theta) = \frac{1}{\theta} e^{-t/\theta}, \qquad \theta > 0,\ t \geq 0.$$

θ is a particular value chosen from $g(\theta)$, and once θ is chosen sampling is done from $f(t|\theta)$. For a sample of size n, the maximum likelihood estimate of θ, $\hat{\theta}_n$ is given by

$$\hat{\theta}_n = \frac{\sum_{i=1}^{n} t_i}{n},$$

where $t_1, t_2, \ldots, t_n$ are the observed times to failure [2].

$f_n(\theta_n|\theta)$, the p.d.f. of $\hat{\theta}_n$ conditioned on θ, is given in [2] as

$$f_n(\hat{\theta}_n|\theta) = \frac{\exp\left(-\frac{n\hat{\theta}_n}{\theta}\right)}{\Gamma(n)} \left(\frac{n}{\theta}\right)^n (\hat{\theta}_n)^{n-1},$$

$$n,\ \theta > 0,\ \hat{\theta}_n \geq 0. \tag{3-1}$$

3.1.1 An Inverted Gamma Prior on θ. Suppose that the prior density on θ, $g(\theta)$, is an inverted gamma. Actually the inverted gamma prior amounts to a gamma prior on the failure rate $(1/\theta)$. The physical reasons for choosing a gamma prior are outlined in [3].

Thus, let

$$g(\theta)=\frac{\alpha^{\lambda}}{\Gamma(\lambda)}\left(\frac{1}{\theta}\right)^{\lambda+1}\exp\left(-\frac{\alpha}{\theta}\right) \qquad \theta\geq 0,\ \alpha,\ \lambda,>0 \tag{3-2}$$

where α and λ are known parameters. Using the theorem on conditional and joint probabilities, it is easy to show that

$$f_n(\hat{\theta}_n)=\int_0^{\infty} f_n(\hat{\theta}_n|\theta)g(\theta)d\theta=\frac{\Gamma(n+\lambda)}{\Gamma(n)\Gamma(\lambda)}\left(\frac{\alpha}{\alpha+n\hat{\theta}_n}\right)^{\lambda}\left(\frac{\hat{n}\theta_n}{\alpha+n\hat{\theta}_n}\right)^{n-1}\left(\frac{n}{\alpha+n\hat{\theta}_n}\right), \tag{3-3}$$

$$0\leq\hat{\theta}_n\leq\infty$$

Using Equations (3–1), (3–2), and (3–3), one can obtain

$$g_n(\theta|\hat{\theta}_n)=\frac{(\alpha+n\hat{\theta}_n)^{n+\lambda}}{\Gamma(n+\lambda)}\left(\frac{1}{\theta}\right)^{n+\lambda+1}\exp\left(-\frac{\alpha+n\hat{\theta}_n}{\theta}\right), \tag{3-4}$$

$$\hat{\theta}_n\geq 0.$$

This is again the inverted gamma density.

Now, the sequence of statistics P_n is

$$P_n=Pr(\theta\geq\theta_1|\hat{\theta}_n)=\int_{\theta_1}^{\infty} g_n(\theta|\hat{\theta}_n)d\theta.$$

Making a change of variable $Z=(\alpha+n\hat{\theta}_n)/\theta$ in Equation (3–4), it follows that

$$P_n-\int_0^{(\alpha+n\hat{\theta}_n)/\theta_1}\frac{1}{\Gamma(n+\lambda)}z^{n+\lambda-1}e^{-z}dz. \tag{3-5}$$

The expression on the right is the incomplete gamma function multiplied by the constant $1/\Gamma(n+\lambda)$. If λ is an integer, Equation (3–5) could for a fixed value of P_n be solved for $\hat{\theta}_n$, by use of either a computer or tables of the Chi-Square distribution. This is because of the fact that 2Z is distributed as a Chi-Square with $2(n+\lambda)$ degrees of freedom. To carry out the test procedure one needs to determine $\hat{\theta}_n^*$ and θ_n* uniquely, and these can be obtained by determining that quantile of the Chi-Square distribution which cuts off an area $1-K_2$, and $1-K_1$ from the right hand tail of the distribution. The θ_n^* and θ_n* are determined by setting $2(\alpha+n\theta_n^*)/\theta_1$ and $2(\alpha+n\theta_n*/\theta_1$ to the corresponding quantile. Ordinarily λ is not an integer. In this case a Chi-Square approximation can be used by rounding off $(n+\lambda)$ to integer values, or by using a computer program that numerically evaluates the quantiles of the gamma distribution for noninteger values of the shape parameter [6].

The Chi Square approximation for noninteger values of λ will be illustrated here.

Let $[u]$ = the largest integer equal to or smaller than u.

Let $\langle u\rangle$ = the smallest integer equal to or larger than u.

Let

$$F_x(\chi^2|v) = \int_0^{\chi^2} \frac{\exp\left(-\frac{x}{2}\right)(x)^{(v/2)-1}}{2^{v/2}\,\Gamma\left(\frac{v}{2}\right)}\,dx \qquad x \geq 0,\ v > 0.$$

That is, $F_x(\chi^2|v)$ is the distribution function of a random variable x which has a Chi-Square distribution with v d.o.f. Since $F_x(\chi^2|v)$ is monotone decreasing in v, one has

$$F_x(\chi^2|v) \leq F_x(\chi^2|\hat{v}), \qquad \text{if } \hat{v} \geq v.$$

In Equation (3–5), put $x = 2Z$, then

$$P_n = F_x\,(\chi^2|2(n+\lambda)\,).$$

Since $2(n+\lambda) \geq 2[n+\lambda]$,

$$P_n = F_x\,(\chi^2|2(n+\lambda)\,) \leq F_x\,(\chi^2|2[n+\lambda]\,) = P_*\ \text{say.}$$

Since $2(n+\lambda) \leq 2\langle n+\lambda\rangle$,

$$P_n = F_x\,(\chi^2|2(n+\lambda)\,) \geq F\,(\chi^2|2\langle n+\lambda\rangle\,) = P^*\ \text{say.}$$

The two critical values $\hat{\theta}_n^*$ and $\hat{\theta}_{n*}$ are found by solving the two equations

$$K_2 = P^* \leq P_n, \text{ and}$$

(3–6).

$$K_1 = P_* \geq P_n.$$

Again, in this example

$$\chi^2 = 2(\alpha + n\hat{\theta}_n)/\theta_1 \text{ and } P_n = P(\theta \geq \theta_1|\hat{\theta}_n).$$

From Equation (3–6) it is evident that whenever $\hat{\theta}_n \epsilon (\hat{\theta}_n^*, \infty)$, $P^* \geq K_2$ and consequently $P_n \geq K_2$, even though due to the approximations being used there may be some $\hat{\theta}_n$'s for which $\hat{\theta}_n \notin (\hat{\theta}_n^*, \infty)$ and yet $P_n \geq K_2$. A similar argument holds for the set $(0, \hat{\theta}_{n*})$.

3.1.2 A Numerical Example The ideas developed in Section 3.1 are illustrated here by a specific numerical example.

Assume a gamma prior and an exponential conditional. Suppose that $K_2 = 0.90$, $K_1 = 0.10$, $\lambda = 0.012$, $\alpha = 2.4$, and $\theta_1 = 100$.

Prior to testing it is necessary to see if $P_0 = Pr(\theta \geq \theta_1)$ lies between K_2 and K_1. Here, since this is the case one proceeds to test as shown below.

Let $n = 1$.

In Equation (3–6), one uses $K_2 = P^* = F(\chi^2|2\langle n+\lambda\rangle) = 0.90$, where $2\langle n+\lambda\rangle = \langle 2.024\rangle = 3$ and $\chi^2 = 2(\alpha + n\hat{\theta}_n^*)/\theta_1$.

But for $P^*=0.9$, and $v=3$, one has $\chi^2=6.2$ [4].

$$\therefore \hat{\theta}_1^*=307.6.$$

Again, in Equation (3–6), one uses $K_1=P_*=F(\chi^2|2[n+\lambda])=0.10$, where $2[n+\lambda]=[2.024]=2$ and $\chi^2=2(\alpha+n\hat{\theta}_n^*)/\theta_1$. But for $P_*=0.1$ and $v=3$, $\chi^2=0.2$ [4].

$$\therefore \hat{\theta}_1^*=7.6.$$

Now, one lifetime is observed; rejection occurs if it fails prior to 7.6 hours, and acceptance occurs if it has not failed by 307.6 hours.

Proceeding in this fashion for $n=2, 3, 4, 5, \ldots$ and so on, Table 1 can be constructed.

TABLE 1

n	$\hat{\theta}_n^*$ Accept if $\hat{\theta}_n \geq \hat{\theta}_n^*$	$\hat{\theta}_{n*}$ Reject if $\hat{\theta}_n \leq \hat{\theta}_n^*$
1	307.6	7.6
2	228.8	26.3
3	199.2	35.9
4	180.6	42.0
5	174.5	47.5

It is to be noted that as n increases, $\hat{\theta}_n^*$ decreases, whereas $\hat{\theta}_{n*}$ increases, implying that the region of indecision narrows down as shown in Figure 1. This is to be expected since the procedure does lead to a decision.

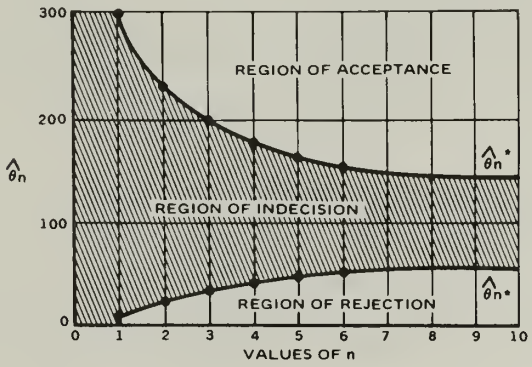


FIGURE 1

3.1.3 A Two-Point Prior on θ Suppose that the prior density on θ, $p(\theta)$, is such that its mass is concentrated at two points θ_A and θ_B only. The reasonableness of such a prior density is discussed in [3].

Let $$Pr(\theta=\theta_A)=p, \text{ and } Pr(\theta=\theta_B)=1-p, \qquad 0<p<1,\ \theta_A<\theta_B.$$

Here again, $P_n=Pr(\theta \geq \theta_1 \mid \hat{\theta}_n)$, and the decision rules would be as follows:

i) if $P_n \geq K_2$, then terminate the test by an acceptance;

ii) if $P_n \leq K_1$, then terminate the test by a rejection; and

iii) if $K_1 < P_n < K_2$, then continue testing.

As before, it may not be necessary to do any testing at all and the following three cases can be identified.

CASE 1:

If $\theta_1 \leq \theta_A < \theta_B$, then accept without testing because $Pr(\theta \geq \theta_1) = 1$.

CASE 2:

If $\theta_A < \theta_B < \theta_1$, then reject without testing because $Pr(\theta \geq \theta_1) = 0$.

CASE 3:

If $\theta_A < \theta_1 \leq \theta_B$, then determine $Pr(\theta \geq \theta_1) = Pr(\theta = \theta_B) = 1 - p$,

if $Pr(\theta \geq \theta_1) \geq K_2$, then accept without testing,

if $Pr(\theta \geq \theta_1) \leq K_1$, then reject without testing, and

if $K_1 < Pr(\theta \geq \theta_1) < K_2$, then conduct the SBP.

$\therefore$ Accept without testing if $1 - p \geq K2$, i.e., if $p \leq 1 - K_2$, and reject without testing if $1 - p \leq K_1$, i.e., if $p \geq 1 - K_1$.

If $K_1 < Pr(\theta \geq \theta_1) < K_2$, then one needs to find

$$P_n = Pr(\theta \geq \theta_1 \mid \hat{\theta}_n) = \sum_{\theta \geq \theta_1} g_n(\theta \mid \hat{\theta}_n).$$

But

$$g_n(\theta|\hat{\theta}_n) = \frac{f_n(\hat{\theta}_n|\theta)p(\theta)}{f_n(\hat{\theta}_n)}, \; f_n(\hat{\theta}_n) > 0,$$

where

$$f_n(\hat{\theta}_n) = \sum_{\text{All } \theta} f_n(\hat{\theta}_n, \theta) = \sum_{\text{All } \theta} f_n(\hat{\theta}_n|\theta)p(\theta)$$

$$= f_n(\hat{\theta}_n|\theta_A)p + f_n(\hat{\theta}_n|\theta_B)\ (1-p).$$

$$\therefore P_n = Pr(\theta \geq \theta_1|\hat{\theta}_n) = \sum_{\theta \geq \theta_1} g_n(\theta|\hat{\theta}_n)$$

$$= \frac{f_n(\hat{\theta}_n|\theta_B)\ (1-p)}{f_n(\theta_n|\theta_A)p + f_n(\hat{\theta}_n|\theta_B)\ (1-p)}.$$

But $f_n(\hat{\theta}_n|\theta)$ is a gamma density, (see Equation (3-1)).

$$\text{Let } E = q\left(\frac{1}{\theta_B}\right)^n \left(e^{-\left(\frac{n\hat{\theta}_n}{\theta_B}\right)}\right)$$

and

$$F = q\left(\frac{1}{\theta_B}\right)^n \left(e^{-\left(\frac{n\hat{\theta}_n}{\theta_B}\right)}\right) + p\left(\frac{1}{\theta_A}\right)^n \left(e^{-\left(\frac{n\hat{\theta}_n}{\theta_A}\right)}\right).$$

Then

$$P_n = E/F,$$

where $q = 1 - p$.

Here again, one can determine a $\hat{\theta}_n^*$ and a $\hat{\theta}_{n*}$ by setting $P_n = K_2$ and $P_n = K_1$, respectively, and proceed with the SBP as illustrated in Section (3.1.2). It is easy to see that the above procedure coincides with a Wald SPRT, and it is easy to verify that the boundaries of the region of indecision do transform into straight lines.

4. A SIMULATION OF SOME SBP LIFE TESTS

To illustrate how the procedure works and the evaluate the goodness of the approximation to $P(A)$ given by Equation 2.3, a computer simulation of some life tests was conducted. This section summarizes this investigation.

An exponential conditional distribution of time to failure with an inverted gamma prior on the scale parameter θ (Section 3.1.1), was the case chosen to be evaluated.

The procedure that was used is explained below:

1. A pair (K_1, K_2) and the minimum acceptable life θ_1 were chosen. A random θ from an inverted gamma distribution with a scale parameter α and a shape parameter λ (integer values only) was then chosen.*

2. Next, a random t, say t_1, from $f(t|\theta)$ was drawn. The test statistic $P_1 = Pr(\theta \geq \theta_1 | \hat{\theta}_1 = t_1)$ was then computed. If $P_1 \geq K_2$ or if $P_1 \leq K_1$, the test was terminated on the first trial resulting in an acceptance or a rejection, respectively.

3. If $K_1 < P_1 < K_2$, a second random t, say t_2, was drawn from $f(t|\theta)$ and the test statistic $P_2 = Pr(\theta \geq \theta_1 | \hat{\theta}_2 = (t_1 + t_2)/2)$ was then computed. The criteria for decision or indecision are the same as those outlined in step 2, and, in general, at the nth step, one has $P_n = P_r(\theta \geq \theta_1 | \hat{\theta}_n)$.

This procedure is repeated for N different draws of θ from the $g(\theta)$. After all the $N\theta$'s have been drawn and the decision to accept or reject has been made for each, an empirical value for the probability of acceptance

$$P(A) = \frac{\text{total number of tests that end in acceptance}}{N}$$

was computed. This $P(A)$ was checked against the approximation to $P(A)$ given in Equation (2–3).

The simulation was repeated for different values of N and different tuples $(K_1, K_2, \alpha, \lambda, \theta_1)$; the results of this are presented in summary form in Table 2. For each case, an empirical frequency distribution of the total number of steps to termination was drawn. A typical distribution (for the case $N = 500$, $\alpha = 20$, $\lambda = 2$, $\theta_1 = 15$, $K_1 = 0.2$, $K_2 = 0.8$) is shown.

The following observations can be made from the results of the simulation:

1. In all 12 cases discussed here, the procedure does lead to a termination. The test terminates more quickly if $P(\theta \geq \theta_1)$ is close to either K_1 or K_2.

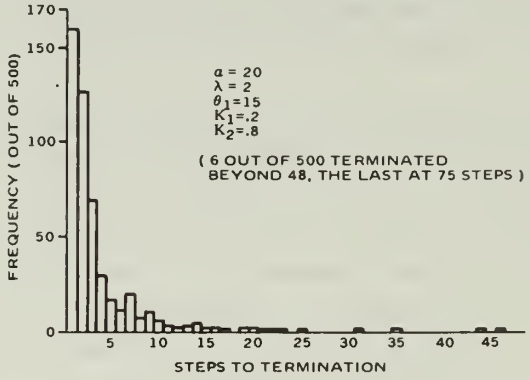
2. The approximation to $P(A)$ is good for tests that take a long time to terminate, i.e., for the cases where $P(\theta \geq \theta_1)$ is not close to either K_1 or K_2.

---

*The tuple $(K_1, K_2, \alpha, \lambda, \theta_1)$ was chosen in such a manner that $K_1 < P(\theta \geq \theta_1) < K_2$. This choice ensured that testing before making a decision was necessary.

TABLE 2.—*Simulation Results*

Sample Size (θ's from $g(\theta)$)	α	λ	θ_1	K_1	K_2	$Pr(\theta \geq \theta_1) = Q$	$P(A) = \frac{Q-K_1}{K_2-K_1}$	Empirical $P(A)$	Sample Mean Number of Steps to Decision	Number of Steps for Which Rejection is Impossible	Step n at Which 99% of Tests Have Terminated
500	20	2	10	0.2	0.8	0.594	0.657	0.602	5.86	1	11
500	20	2	10	0.15	0.85	0.594	0.634	0.624	10.02	2	19
500	20	2	10	0.1	0.9	0.594	0.617	0.560	15.48	2	33
500	20	2	10	0.05	0.95	0.594	0.604	0.654	32.87	3	148
500	20	2	15	0.2	0.8	0.385	0.308	0.318	4.32	0	8
500	20	2	15	0.15	0.85	0.385	0.336	0.324	7.34	1	12
500	20	2	15	0.1	0.9	0.385	0.356	0.354	14.93	1	29
500	20	2	15	0.05	0.95	0.385	0.372	0.414	27.12	1	64
2000	300	3	100	0.5	0.9	0.577	0.192	0.298	2.91	0	6
2000	300	3	100	0.1	0.65	0.577	0.867	0.708	7.81	2	16
2000	300	3	100	0.2	0.8	0.577	0.628	0.576	8.44	1	18
2000	300	3	100	0.1	0.9	0.577	0.596	0.597	22.53	3	57



FIGURE

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge the efforts of Dr. James Lechner of Research Analysis Corporation, McLean, Virginia, and Mr. Jacob Bricker for their comments on a previous draft of this paper. Mr. John Collins did the necessary programming for the simulations and other numerical work.

REFERENCES

[1] Bricker, J. L. and R. E. Schafer, "A Sequential Bayes Test of a Composite Hypothesis," submitted for publication to Journal of the American Statistical Association.

[2] Epstein, B. and M. Sobel, "Some Theorems Relevent to Life Testing From an Exponential Distribution," Annals of Mathematical Statistics *27*, 15–25 (1956).

[3] Harris, C. and N. Singpurwalla, "Life Distributions Derived From Stochastic Hazard Functions," IEEE Transaction On Reliability Vol. R–17, No. 2 (June 1968).

[4] Pearson, E. S. and H. O. Hartley, Biometrika Tables for Statisticians (Cambridge University Press, Cambridge, Mass., 1966), 3rd ed., Vol. 1.

[5] Schafer, R. E., "Bayes Single Sampling Plans by Attributes Based on Posterior Risks," Naval Research Logistics Quarterly *14*, 81–88 (Mar. 1967).

[6] Wilk, M. B., R. Gnanadesikan, and M. Huyett, "Probability Plots for the Gamma Distribution," Technometrics *4*, 1–20 (1962).

* * *

CONFIDENCE INTERVALS FOR RANKED MEANS*

Edward J. Dudewicz

The University of Rochester

ABSTRACT

Suppose that observations from populations $\pi_1, \ldots, \pi_k$ $(k \geq 1)$ are normally distributed with unknown means $\mu_1, \ldots, \mu_k$, respectively, and a common known variance σ^2. Let $\mu_{[1]} \leqq \ldots \leqq \mu_{[k]}$ denote the ranked means. We take n independent observations from each population, denote the sample mean of the n observation from π_i by $\bar{X}_i$ $(i=1, \ldots, k)$, and define the ranked sample means $\bar{X}_{[1]} \leqq \ldots \leqq \bar{X}_{[k]}$. The problem of confidence interval estimation of $\mu_{[1]}, \ldots, \mu_{[k]}$ is stated and related to previous work (Section 1). The following results are obtained (Section 2). For $i=1, \ldots, k$ and any $\gamma (0<\gamma<1)$ an upper confidence interval for $\mu_{[i]}$ with minimal probability of coverage γ is $(-\infty, \bar{X}_{[i]}+h_i^*)$ with $h_i^* = (\sigma/n^{1/2})\Phi^{-1}(\gamma^{1/k-i+1})$, where $\Phi(\cdot)$ is the standard normal cdf. A lower confidence interval for $\mu_{[i]}$ with minimal probability of coverage γ is $(\bar{X}_{[i]}-g_i^*, +\infty)$ with $g_i^* = (\sigma/n^{1/2})\,\Phi^{-1}(\gamma^{1/i})$. For the upper confidence interval on $\mu_{[i]}$ the maximal probability of coverage is $1-[1-\gamma^{1/k-i+1}]^i$, while for the lower confidence interval on $\mu_{[i]}$ the maximal probability of coverage is $1-[1-\gamma^{1/i}]^{k-i+1}$. Thus the maximal overprotection can always be calculated. The overprotection is tabled for $k=2, 3$. These results extend to certain translation parameter families.It is proven that, under a bounded completeness condition, a monotone upper confidence interval $h(\bar{X}_1, \ldots, \bar{X}_k)$ for $\mu_{[i]}$, with probability of coverage $\gamma (0<\gamma<1)$ for all $\mu = (\mu_{[1]}, \ldots, \mu_{[k]})$, does not exist.

1. INTRODUCTION

Consider $k(\geq 2)$ populations $\pi_1, \ldots, \pi_k$ such that observations from π_i are normal $N(\mu_i, \sigma^2)$ with unknown mean μ_i and common known variance $\sigma_i^2 = \sigma^2 (i=1, \ldots, k)$, and the rule: Take n independent vectors $X_j = (X_{1j}, \ldots, X_{kj})$, $j=1, \ldots, n$, where X_{ij} denotes the jth observation from π_i; form the sample means $\bar{X}_i = \sum_{j=1}^{n} X_{ij}/n$ $(i=1, \ldots, k)$, and base the terminal decision solely on them. (This rule has been utilized in many instances of statistical decision problems. For example, several "ranking and selection" procedures (e.g., that of Gupta (1965)) use this rule.) Let $\mu_{[1]} \leqq \ldots \leqq \mu_{[k]}$ denote the ranked values of $\mu_1, \ldots, \mu_k$. We assume throughout that both $\mu_1, \ldots, \mu_k$ and the pairings of $\mu_{[1]}, \ldots, \mu_{[k]}$ with $\pi_1, \ldots, \pi_k$ are completely unknown (although we vary the distribution from normality) and consider the problem: estimate $\mu_{[i]} (1 \leqq i \leqq k)$ based on $\bar{X}_1, \ldots, \bar{X}_k$.

Let $\mathscr{R} = \{x: -\infty < x < \infty\}$. Let $\Omega_o(a) = \{(\mu_1, \ldots, \mu_k): \mu_i \in \mathscr{R} (i=1, \ldots, k),\ a \text{ is held fixed}\}$. (In general, a will be one of $\mu_{[1]}, \ldots, \mu_{[k]}$.) Let $\bar{X}_{[1]} \leqq \ldots \leqq \bar{X}_{[k]}$ denote the ordered $\bar{X}_i (i=1, \ldots, k)$. (We disregard the possibility of ties, which occur w.p. 0 below.) Denote the $N(0, 1)$ distribution function (d.f.) and density function (fr.f.) by $\Phi(\cdot)$ and $\phi(\cdot)$, respectively. Let S_k be the symmetric group on k elements.

---

*This research was supported by Office of Naval Research contracts N00014-68A-0091, and Nonr 401(53), and by NSF grants GP-8958 and GP-7798. "This research is supported in whole or in part by The Center for Naval Analyses of the University of Rochester. Such support does not imply endorsement of the content by The Navy."

(1) DEFINITION: Let F and f be the respective d.f. and fr.f. of observations from an arbitrary univariate location parameter family; i.e.,

$$F(x;\ \theta)=F(x-\theta)\ (x\epsilon\mathscr{R}),$$

where $F(\cdot)$ is a d.f. with fr.f. $f(\cdot)$ and θ is fixed, $\theta\epsilon\Theta\subseteq\mathscr{R}$.

Consider joint confidence interval estimation of $\mu_{[1]}, \ldots, \mu_{[k]}$. We observe X_{ij} $(i=1, \ldots, k;\ j=1, \ldots, n)$, take $\bar{X}_1, \ldots, \bar{X}_k$ to be fundamental (note that $\bar{X}_1, \ldots, \bar{X}_k$ are sufficient and transitive for $\mu_1, \ldots, \mu_k$ after n stages), choose interval form

$$(2)\qquad I=I(\bar{X}_1, \ldots, \bar{X}_k)=\{\mu_{[1]}, \ldots, \mu_{[k]}: g_1 \leqslant \mu_{[1]} \leqslant h_1, \ldots, g_k \leqslant \mu_{[k]} \leqslant h_k\},$$

where $g_1, h_1; \ldots; g_k, h_k$ are functions of $\bar{X}_1, \ldots, \bar{X}_k$, and ask two invariance conditions (involving relabeling of populations and shifts of location):

$$(3)\qquad I(\bar{X}_1, \ldots, \bar{X}_k)=I(\bar{X}_{\beta(1)}, \ldots, \bar{X}_{\beta(k)}) \text{ for all } B\epsilon S_k;$$

$$(4)\qquad I(\bar{X}_1+c, \ldots, \bar{X}_k+c)=I(\bar{X}_1, \ldots, \bar{X}_k)+c \text{ for all } c\epsilon\mathscr{R}.$$

(5) DEFINITION: Let $a_1, \ldots, a_k$ $(a_1 \geqslant 0, \ldots, a_k \geqslant 0;\ a_1+\ldots+a_k=1)$, b^* $(0<b^*<\infty)$, and (G, H) $(-\infty \leqslant G \leqslant H \leqslant +\infty)$ be constants preset by the experimenter.

Now take as loss function a weighted sum of the probability that $I(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]})$ doesn't cover $\mu_{[i]}$ plus a multiple of a quantity related to the length of the interval on $\mu_{[i]}$ $(i=1, \ldots, k)$:

$$(6)\qquad W(\mu;\ I(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]}))=\sum_{i=1}^{k} a_i\{P_\mu[\mu_{[i]} \notin (g_i, h_i)]+b^* \min\ (h_i-g_i, H-G)\}.$$

(Note that the length h_i-g_i is the special case of min $(h_i-g_i, H-G)$ where the experimenter chooses (G, H) with $H-G=+\infty$.) The risk function

$$(7)\qquad r(\mu;\ I(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]}))=E_\mu W(\mu;\ I(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]}))$$

is then

$$(8)\qquad r(\mu;\ I)=\sum_{i=1}^{k} a_i E_\mu\{P_\mu[\mu_{[i]} \notin (g_i, h_i)]+b^* \min\ (h_i-g_i, H-G)\}.$$

Functions $g_1, h_1; \ldots; g_k, h_k$ which achieve

$$(9)\qquad \inf_{g_1, h_1; \ldots; g_k, h_k}\ \sup_{\mu\epsilon\Omega_0(\mu_{[k]})}\ r(\mu;\ I(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]}))$$

and provide a minimax invariant confidence interval are not now available. Note that by (3) and (4) with $c=-\bar{X}_{[i]}$ we have the

(10) THEOREM: Under conditions (3) and (4) $I(\bar{X}_1, \ldots, \bar{X}_k)$ must be of the form $I(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]})$ with (for $i=1, \ldots, k$)

$$g_i = \bar{X}_{[i]} - g_i^*(\bar{X}_{[i]} - \bar{X}_{[1]}, \ldots, \bar{X}_{[i]} - \bar{X}_{[k]})$$

$$h_i = \bar{X}_{[i]} + h_i^*(\bar{X}_{[i]} - \bar{X}_{[1]}, \ldots, \bar{X}_{[i]} - \bar{X}_{[k]}).$$

The subclass of joint intervals $I_N(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]})$ utilizes the "natural" estimators $\bar{X}_{[i]}$ of $\mu_{[i]}$ $(i=1, \ldots, k)$ strongly by taking $g_1^*, h_1^*; \ldots; g_k^*, h_k^*$ (Theorem (10)) to be constants.

Previous work. Further suppose the experimenter has specified positive constants $d_1, \ldots, d_k$, and wishes the interval about $\mu_{[i]}$ to be of length d_i $(i=1, \ldots, k)$. The intervals of fixed width within I_N comprise the subclass

$$(11) \qquad I_{F,N}(\bar{X}_{[1]}, \ldots, \bar{X}_{[k]}) = \{\mu_{[1]}, \ldots, \mu_{[k]} : \bar{X}_{[i]} + (h_i^* - d_i) \leq \mu_{[i]} \leq \bar{X}_{[i]} + h_i^* \ (i=1, \ldots, k)\}.$$

Then (here it is logical to choose $(G, H) = (-\infty, +\infty)$)

$$(12) \qquad r(\mu; I_{F,N}) = 1 + b^* \sum_{i=1}^{k} a_i d_i - \sum_{i=1}^{k} a_i P_\mu[\mu_{[i]} - h_i^* \leq \bar{X}_{[i]} \leq \mu_{[i]} - h_i^* + d_i].$$

For the case $a_1 = \ldots = a_{k-1} = 0$, $a_k = 1$, set $h_k^* = d_k/2$. Then Lal Saxena and Tong [5] prove that

$$(13) \qquad \inf_{\mu \in \Omega_o(\mu_{[k]})} P_\mu[\mu_{[k]} - d_k/2 \leq \bar{X}_{[k]} \leq \mu_{[k]} + d_k/2],$$

occurs at $\mu_{[1]} = \ldots = \mu_{[k]}$, and therefore equals

$$\left[\Phi\left(\frac{d_k}{2}\frac{\sqrt{n}}{\sigma}\right)\right]^k - \left[\Phi\left(-\frac{d_k}{2}\frac{\sqrt{n}}{\sigma}\right)\right]^k;$$

i.e., the interval $(\bar{X}_{[k]} - d_k/2, \bar{X}_{[k]} + d_k/2)$ for $\mu_{[k]}$ has minimum probability of coverage when

$$\mu_{[1]} = \ldots = \mu_{[k]}.$$

2. UPPER AND LOWER INTERVALS; NONEXISTENCE

The subclass of joint intervals I_N utilizes $\bar{X}_{[i]}$ as an estimator of $\mu_{[i]}$ strongly $(i=1, \ldots, k)$. For problems in which we wish an upper (lower) joint confidence interval on $\mu_{[1]}, \ldots, \mu_{[k]}$, we will set $g_1^* = \ldots = g_k^* = +\infty$ $(h_1^* = \ldots = h_k^* = +\infty)$. Then our interval is in one of the classes

$$(14) \qquad I_{N,U} = \{\mu_{[1]}, \ldots, \mu_{[k]} : \mu_{[1]} \leq \bar{X}_{[1]} + h_1^*, \ldots, \mu_{[k]} \leq \bar{X}_{[k]} + h_k^*\}$$

$$(15) \qquad I_{N,L} = \{\mu_{[1]}, \ldots, \mu_{[k]} : X_{[1]} - g_1^* \leq \mu_{[1]}, \ldots, \bar{X}_{[k]} - g_k^* \leq \mu_{[k]}\}$$

and

$$(16) \qquad r(\mu; I_{N,U}) = 1 - \sum_{i=1}^{k} a_i P_\mu[\bar{X}_{[i]} \geq \mu_{[i]} - h_i^*] + b^*(H - G)$$

$$(17) \qquad r(\mu; I_{N,L}) = 1 - \sum_{i=1}^{k} a_i P_\mu[\bar{X}_{[i]} \leq \mu_{[i]} + g_i^*] + b^*(H - G).$$

For the case of upper intervals we may choose $H-G=0$ without loss. Then

$$r(\mu;\, I_{N,U}) = \sum_{i=1}^{k} a_i P_\mu[\bar{X}_{[i]} \leq \mu_{[i]} - h_i^*]. \tag{18}$$

Similarly, for the case of lower intervals we may choose $H-G=0$ without loss. Then

$$r(\mu;\, I_{N,L}) = \sum_{i=1}^{k} a_i P_\mu[\bar{X}_{[i]} \geq \mu_{[i]} + g_i^*]. \tag{19}$$

(20) THEOREM: For any $i(1 \leq i \leq k)$, if $a_i=1$ then the risk (18) ((19)) is the probability that our upper (lower) interval doesn't cover $\mu_{[i]}$, and is maximized over $\mu \epsilon \Omega_o(\mu_{[i]})$ at

$$\mu = \left(\underbrace{-\infty, \ldots, -\infty}_{i-1 \text{ terms}}, \underbrace{\mu_{[i]}, \ldots, \mu_{[i]}}_{k-i+1 \text{ terms}} \right) \left(\mu = \left(\underbrace{\mu_{[i]}, \ldots, \mu_{[i]}}_{i \text{ terms}}, \underbrace{+\infty, \ldots, +\infty}_{k-i \text{ terms}} \right) \right).$$

Thus, for any $\gamma(0<\gamma<1)$ an upper (lower) confidence interval of minimal probability of coverage γ is $(-\infty, \bar{X}_{[i]} + h_i^*)$ $((\bar{X}_{[i]} - g_i^*, +\infty))$ with $h_i^* = (\sigma/\sqrt{n})\Phi^{-1}(\gamma^{1/k-i+1})$ $(g_i^* = (\sigma/\sqrt{n})\Phi^{-1}(\gamma^{1/i}))$.

PROOF: *Upper Interval.* For any $i(1 \leq i \leq k)$, if $a_i=1$, then $a_j=0 (j \neq i)$ (see (5)) and

$$\sup_{\mu \epsilon \Omega_o(\mu_{[i]})} r(\mu;\, I_{N,U}) = \sup_{\mu \epsilon \Omega_o(\mu_{[i]})} P_\mu[\bar{X}_{[i]} \leq \mu_{[i]} - h_i^*] = \sup_{\mu \epsilon \Omega_o(\mu_{[i]})} F_{\bar{X}_{[i]}}(\mu_{[i]} - h_i^*)$$

$$= \lim_{M \to +\infty} P_{\mu_{[1]} = \ldots = \mu_{[i-1]} = -M,\, \mu_{[i]} = \ldots = \mu_{[k]}}[\bar{X}_{[i]} \leq \mu_{[i]} - h_i^*]$$

since, for $i=1, \ldots, k$ and $x \epsilon \mathscr{R}$, $F_{\bar{X}_{[i]}}(x) \downarrow$ as $\mu_l \uparrow$ $(l=1, \ldots, k)$ by Theorem $(A\text{–}1)$ (see Appendix). It follows that

$$\sup_{\mu \epsilon \Omega_o(\mu_{[i]})} r(\mu;\, I_{N,U}) = P_{\mu_{[1]} = \ldots = \mu_{[i-1]} = -\infty,\, \mu_{[i]} = \ldots = \mu_{[k]}}[\bar{X}_{[i]} \leq \mu_{[i]} - h_i^*]$$

$$= P\left[\min\,(Y_1, \ldots, Y_{k-i+1}) \leq \frac{-h_i^*}{\sigma/\sqrt{n}} \right],$$

where $Y_1, \ldots, Y_{k-i+1}$ are $k-i+1$ independent $N(0, 1)$ r.v.'s. To make the minimal probability of coverage $\gamma(0<\gamma<1)$, we set h^* so that

$$1-\gamma = 1 - \left[1 - \Phi\left(\frac{-h_i^*}{\sigma/\sqrt{n}} \right) \right]^{k-i+1} = 1 - \left[\Phi\left(\frac{h_i^*}{\sigma/\sqrt{n}} \right) \right]^{k-i+1};$$

thus

$$\gamma = \left[\Phi\left(\frac{h_i^*}{\sigma/\sqrt{n}} \right) \right]^{k-i+1},\ \gamma^{1/k-i+1} = \Phi\left(\frac{h_i^*}{\sigma/\sqrt{n}} \right),\ \text{and}\ h_i^* = (\sigma/\sqrt{n})\Phi^{-1}(\gamma^{1/k-i+1}).$$

Lower Interval. This case follows in a similar manner.

(21) THEOREM: The upper confidence interval of (20) on $\mu_{[i]}$ which has minimal probability of coverage γ has maximal probability of coverage $1-[1-\gamma^{1/k-i+1}]^i$ $(i=1, \ldots, k;\ 0<\gamma<1)$.

The lower confidence interval of (20) on $\mu_{[i]}$ which has minimal probability of coverage γ has maximal probability of coverage $1-[1-\gamma^{1/i}]^{k-i+1}$ $(i=1, \ldots, k;\ 0<\gamma<1)$.

The proof of Theorem (21) is similar to that of Theorem (20) and will be omitted. Note that (20) and (21) also hold when $k=1$, in which case the upper and lower intervals on $\mu_{[1]}$ are exact. Table 1 illustrates the maximal degree of overprotection.

(22) TABLE 1.

$1-[1-\gamma^{1/k-i+1}]^i$

k	1	2		3		
γ	$i=1$	$i=1$	$i=2$	$i=1$	$i=2$	$i=3$
0.99	0.99	0.995	1.000	0.997	1.000	1.000
0.95	0.95	0.975	0.998	0.983	0.999	1.000
0.90	0.90	0.949	0.99	0.965	0.997	0.999
0.80	0.80	0.894	0.96	0.923	0.989	0.992
0.70	0.70	0.837	0.91	0.888	0.973	0.973
0.60	0.60	0.775	0.84	0.843	0.949	0.936
0.50	0.50	0.707	0.75	0.794	0.914	0.875

For the special case $i=k$, Fraser [3], p. 579, gave the upper interval on $\mu_{[k]}$ of Theorem (20) as one with probability of coverage *at least* γ. Fraser proves that under mild conditions an upper confidence interval for $\mu_{[k]}(k \geq 2)$, with probability of coverage $\gamma(0<\gamma<1)$ for all $\mu \epsilon \Omega_o$, does not exist.

Our results above extend to certain location parameter families if, instead of normal distributions we take (23): a location parameter family with finite mean, i.e.:

(23) Given $k(\geq 2)$ populations $\pi_1, \ldots, \pi_k$ such that observations from population π_i have fr.f. $f(x-\theta_i)$, $x \epsilon \mathscr{R}$, where the location parameter θ_i is unknown $(i=1, \ldots, k)$. Assume the fr.f. f is such that $E_f \equiv \int_{-\infty}^{\infty} x f(x)dx < \infty$.

Then we may talk of $\mu_1, \ldots, \mu_k$ (or of $\mu_{[1]}, \ldots, \mu_{[k]}$). Denote the ranked values of the location parameters $\theta_1, \ldots, \theta_k$ by $\theta_{[1]} \leq \ldots \leq \theta_{[k]}$.

Then since

$$\int_{-\infty}^{\infty} x f(x-\theta)dx = \int_{-\infty}^{\infty} x f(x)dx + \theta = E_f + \theta,$$

the population associated with $\mu_{[i]}$ is precisely the population associated with $\theta_{[i]}$ $(i=1, \ldots, k)$. It is well-known that the distribution of $X_{ij}-\theta_i$ does not depend on θ_i $(i=1, \ldots, k)$. Define

$$G_n(y|f) = P[\{(X_{i1}-\theta_i)+ \ldots +(X_{in}-\theta_i)\}/n \leq y], \qquad g_n(y|f) = \frac{d}{dy} G_n(y|f).$$

Then for $i=1, \ldots, k$,

$$P[\bar{X}_i \leq x] = P[X_{i1}+ \ldots +X_{in} \leq nx] = P[(X_{i1}-\theta_i)+ \ldots +(X_{in}-\theta_i) \leq n(x-\theta_i)] = G_n(x-\theta_i|f).$$

(24) THEOREM: Suppose we have location parameter populations as in (23). For any i $(1 \leq i \leq k)$, if $a_i=1$ (thus $a_j=0$ for $j \neq i$) then the risk (18) ((19)) is the probability that our upper (lower) interval does not cover $\mu_{[i]}$ and is maximized over $\mu \epsilon \Omega_o(\mu_{[i]})$ at

$$\mu=\Big(\underbrace{-\infty, \ldots, -\infty}_{i-1 \text{ terms}}, \underbrace{\mu_{[i]}, \ldots, \mu_{[i]}}_{k-i+1 \text{ terms}}\Big)\Big(\mu=\Big(\underbrace{\mu_{[i]}, \ldots, \mu_{[i]}}_{i \text{ terms}}, \underbrace{+\infty, \ldots, +\infty}_{k-i \text{ terms}}\Big)\Big).$$

Thus, for any γ $(0<\gamma<1)$ an upper (lower) confidence interval of minimal probability of coverage γ is $(-\infty, \bar{X}_{[i]}+h_i^*)$ $((\bar{X}_{[i]}-g_i^*, +\infty))$ with $h_i^*=-G_n^{-1}$ $(1-\gamma^{1/(k-i+1)}|f)+E_f$

$$(g_i^*=G_n^{-1}\ (\gamma^{1/i}|f)-Ef).$$

If $g_n(x|f)$ is symmetric about $x=0$ this becomes $h_i^*=G_n^{-1}$ $(\gamma^{1/(k-i+1)}|f)+E_f(g_i^*=G_n^{-1}(\gamma^{1/i}|f)-E_f)$.

PROOF: *Upper Interval.* For any i $(1\leq i\leq k)$, if $a_i=1$, $a_j=0$ $(j\neq i)$ then by Theorem (A–1)

$$\begin{aligned}\sup_{\mu\epsilon\Omega_0(\mu_{[i]})} r(\mu; I_{N,U}) &= \sup_{\mu\epsilon\Omega_0(\mu_{[i]})} P_\mu[\bar{X}_{[i]}\leq \mu_{[i]}-h_i^*]\\ &=\lim_{M\to+\infty} P_{\mu_{[1]}=\ldots=\mu_{[i-1]}=-M,\,\mu_{[i]}=\ldots=\mu_{[k]}}[\bar{X}_{[i]}\leq \mu_{[i]}-h_i^*]\\ &=\lim_{M\to+\infty} H_M(\mu_{[i]}-h_i^*),\end{aligned}$$

where $H_M(x)=P_\mu[\bar{X}_{[i]}\leq x]$ with $\mu=(-M, \ldots, -M, \mu_{[i]}, \ldots, \mu_{[i]})$. Now $H_M(x)\to H_\infty(x)$ for all x by the expression for $F_{\bar{X}_{[i]}}(x)$ given in Lemma (A–2) (see appendix). Thus

$$\begin{aligned}\sup_{\mu\epsilon\Omega_0(\mu_{[i]})} r(\mu; I_{N,U}) &= P_{\mu_{[1]}=\ldots=\mu_{[i-1]}=-\infty,\,\mu_{[i]}=\ldots=\mu_{[k]}}[\bar{X}_{[i]}\leq \mu_{[i]}-h_i^*]\\ &=P[\min(Y_1, \ldots, Y_{k-i+1})\leq -h_i^*+E_f],\end{aligned}$$

where $Y_1, \ldots, Y_{k-i+1}$ are $k-i+1$ independent r.v.'s each with d.f $G_n(y|f)$. It follows that to make the minimal probability of coverage $\gamma(0<\gamma<1)$ we set h_i^* so that $h_i^*=-G_n^{-1}(1-\gamma^{1/k-i+1}|f)+E_f$.

Lower Interval. This case follows in a similar manner.

(25) THEOREM: The upper confidence interval of (24) on $\mu_{[i]}$ which has minimal probability of coverage γ has maximal probability of coverage $1-[1-g^{1/k-i+1}]^2$ $(i=1, \ldots, k;\ 0<\gamma<1)$.

The lower confidence interval of (24) on $\mu_{[i]}$ which has minimal probability of coverage γ has maximal probability of coverage $1-[1-\gamma^{1/i}]^{k-i+1}$ $(i=1, \ldots, k;\ 0<\gamma<1)$.

The proof of Theorem (25) will be omitted. Note that this result implies that Table 1 provides an analysis of maximal overprotection for our location parameter case as well as for the normal case. For the special case $i=k$, Fraser [3], p. 576, gave the upper interval on $\mu_{[k]}$ of Theorem (24) as one with probability of coverage *at least* γ. Fraser proves that under mild conditions an upper confidence interval for $\mu_{[k]}$ $(k\geq 2)$, with probability of coverage γ $(0<\gamma<1)$ for all $\mu\epsilon\Omega_o$, does not exist if $f(x-\mu)$ satisfies a condition of bounded completeness. We will now extend this result to $\mu_{[i]}$ $(1\leq i\leq k;\ k\geq 2)$; our mild conditions are slightly stronger than Fraser's.

(26) DEFINITION: Let $h(x_1, \ldots, x_k)$ be a real-valued function such that for any j $(1\leq j\leq k)$

$$h(x_1, \ldots, x_k)\leq h(x_1, \ldots, x_{j-1}, x_j+\delta, x_{j+1}, \ldots, x_k) \text{ for all } x_1, \ldots, x_k\epsilon\mathscr{R} \text{ and } \delta>0.$$

(27) DEFINITION: Let

$$\phi_\theta(x_1, \ldots, x_k)=\begin{cases}1 \text{ if } h(x_1, \ldots, x_k)\geq\theta\\ 0 \text{ if } h(x_1, \ldots, x_k)<\theta.\end{cases}$$

(28) DEFINITION: For any i $(1\leq i\leq k)$ for $l=1, 2, \ldots$ let

$$R_i(y_1, \ldots, y_l) = \begin{cases} \text{the } i\text{th smallest of } y_1, \ldots, y_l & \text{if } l \geq i \\ +\infty & \text{if } l < i. \end{cases}$$

Let $R_o(y_1, \ldots, y_l) = -\infty$ if $l \geq 1$. Let $R_i(x_j) \equiv R_i(x_1, \ldots, x_k)$.

(29) DEFINITION: For $1 \leq l \leq k$, let

$$S_l(i) \equiv S_l = \{(x_1, \ldots, x_k) : R_i(x_j, j \neq l) > x_1 > R_{i-1}(x_j, j \neq l)\}.$$

Note that $\phi_\theta(x_1, \ldots, x_k)$ is a monotone non-decreasing function of $x_1, \ldots, x_k$ and that $S_1, \ldots, S_k$ are disjoint sets whose union is $\mathcal{R}^k$.

(30) THEOREM: Suppose we have location parameter populations as in (23) and $G_n(y-\theta|f)$ is boundedly complete (each-sided), i.e. $Eg(X) = \int_{-\infty}^{\infty} g(x)\,dG_n(x-\theta|f) = 0$ for a dense set of $\theta (< 0 \text{ or } > 0)$ and $|g(x)| < M$ imply $g(x) = 0$ (a.e.). Fix i $(1 \leq i \leq k;\ k \geq 2)$. Then an upper confidence interval for $\mu_{[i]}$, with probability of covering γ $(0 < \gamma < 1)$ for all $\mu \epsilon \Omega_o$, and satisfying (26) does not exist.

PROOF: (We in general follow Fraser [3].) Assume that $h(x_1, \ldots, x_k)$ satisfies (26) and yields an upper confidence interval for $\mu_{[i]}$ with probability of covering $\gamma (0 < \gamma < 1)$ for all $\mu \epsilon \Omega_o$. We have

$$\gamma \overset{\mu}{=} P_\mu[\mu_{[i]} \leq h(\bar{X}_1, \ldots, \bar{X}_k)] = E\phi_{\mu_{[i]}}(\bar{X}_1, \ldots, \bar{X}_k) \tag{31}$$

$$= E\int_{-\infty}^{\infty} \phi_{\mu_{[i]}}(\bar{X}_1, \ldots, \bar{X}_{l-1}, x_l, \bar{X}_{l+1}, \ldots, \bar{X}_k)\,dG_n(x_l - \mu_{[i]} + E_f|f) \text{ if } \mu_l = \mu_{[i]}$$

$$= E[\beta^{(l)}_{\mu_{[i]}}(\bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k)] \text{ if } \mu_l = \mu_{[i]},$$

where

$$\beta^{(l)}_{\mu_{[i]}}(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) = \int_{-\infty}^{\infty} \phi_{\mu_{[i]}}(x_1, \ldots, x_k)\,dG_n(x_l - \mu_{[i]} + E_f|f).$$

We now derive conditions on the function $\beta^{(l)}_o$. From (31) it follows that (if $\mu_l = \mu_{[i]}$)

$$E[\beta^{(l)}_o(\bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k) - \gamma] = 0.$$

Hence, by bounded completeness,

$$\beta^{(l)}_o(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) = \gamma \text{ (a.e.)}.$$

Using the above condition on $\beta^{(l)}_o$, we obtain conditions on the function $\phi_o(x_1, \ldots, x_k)$.

$$\gamma = \beta^{(l)}_o(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) \text{ (a.e.)}$$

$$= \int_{-\infty}^{\infty} \phi_o(x_1, \ldots, x_k)\,dG_n(x_l + E_f|f).$$

Consider fixed $x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k$. Now $\phi_o(x_1, \ldots, x_k)$ is a monotone function of x_l, and

since it is a characteristic function it will have the following form:

$$u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) = \max \left[\begin{array}{l} \text{value of } x_l \text{ at which} \\ \phi_0(x_1, \ldots, x_k) \text{ jumps, } R_{i-1}(x_j, j \neq l) \\ \text{from 0 to 1} \end{array} \right]$$

$$\phi_0(x_1, \ldots, x_k) = \begin{cases} 0 \text{ if } R_{i-1}(x_j, j \neq l) < x_l < u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k). \\ 1 \text{ if } u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) < x_l < \infty. \end{cases}$$

Using the function $u(x_1, \ldots, x_k)$, we obtain

$$\gamma = \int_{-\infty}^{R_{i-1}(x_j, j \neq l)} \phi_0(x_1, \ldots, x_k) dG_n(x_l + E_f | f) + \int_{u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k)}^{\infty} dG_n(x_l + E_f | f). \tag{32}$$

However, since

$$0 \leq \int_{-\infty}^{R_{i-1}(x_j, j \neq l)} \phi_0(x_1, \ldots, x_k) dG_n(x_l + E_f | f)$$

$$\leq \int_{-\infty}^{R_{i-1}(x_j, j \neq l)} dG_n(x_l + E_f | f) = P[\bar{X}_l \leq R_{i-1}(x_j, j \neq l) + E_f],$$

then

$$G_n^{-1}(1 - \gamma | f) - E_f \leq u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k).$$

The inequality on $u(x_1, \ldots, x_k)$ implies that $\phi_0(x_1, \ldots, x_k) = 0$ for $(x_1, \ldots, x_k) \epsilon S_l$ with $x_l < G_n^{-1}(1 - \gamma | f) - E_f$. This is true for $l = 1, \ldots, k$; hence

$$\phi_0(x_1, \ldots, x_k) = 0 \text{ if } R_i(x_j) < G_n^{-1}(1 - \gamma | f) - E_f.$$

Consider now $(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k)$ having $R_{i-1}(x_j, j \neq l) < G_n^{-1}(1 - \gamma | f) - E_f$; in (32) the first integral vanishes leaving

$$\gamma = \int_{u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k)}^{\infty} dG_n(x_l + E_f | f).$$

Therefore $u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k) = G_n^{-1}(1 - \gamma | f) - E_f$, if $R_{i-1}(x_j, j \neq l) < G_n^{-1}(1 - \gamma | f) - E_f$. From this equality on $u(x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_k)$, we obtain the following conditions on $\phi_0(x_1, \ldots, x_k)$:

$$\phi_0(x_1, \ldots, x_k) = \begin{cases} 0 \text{ if } R_i(x_j) < G_n^{-1}(1 - \gamma | f) - E_f \\ 1 \text{ if } \begin{cases} R_i(x_j) > G_n^{-1}(1 - \gamma | f) - E_f \\ R_{i-1}(x_j) < G_n^{-1}(1 - \gamma | f) - E_f. \end{cases} \end{cases}$$

But since $\phi_0(x_1 + \delta, \ldots, x_k + \delta)$ is monotone in δ, we have

$$\phi_0(x_1, \ldots, x_k) = \begin{cases} 0 \text{ if } R_i(x_j) < G_n^{-1}(1 - \gamma | f) - E_f \\ 1 \text{ if } R_i(x_j) > G_n^{-1}(1 - \gamma | f) - E_f. \end{cases}$$

Therefore,

$$h(x_1, \ldots, x_k) \begin{matrix} < 0 \text{ if } R_i(x_j) < G_n^{-1}(1-\gamma|f) - E_f \\ \geq 0 \text{ if } R_i(x_j) > G_n^{-1}(1-\gamma|f) - E_f. \end{matrix}$$

Similarly

$$h(x_1, \ldots, x_k) \begin{matrix} < \mu_{[i]} \text{ if } R_i(x_j) < G_n^{-1}(1-\gamma|f) + \mu_{[i]} - E_f \\ \geq \mu_{[i]} \text{ if } R_i(x_j) > G_n^{-1}(1-\gamma|f) + \mu_{[i]} - E_f \end{matrix}$$

This completely determines $h(x_1, \ldots, x_k)$: $h(x_1, \ldots, x_k) = R_i(x_j) - G_n^{-1}(1-\gamma|f) + E_f$. But we know that a constant added to this yields $X_{[i]} - G_n^{-1}(1-\gamma^{1/(k-i+1)}|f) + E_f$, which doesn't *always* yield γ; therefore this can't. (Note: it does yield γ for *some* μ vectors.)

Note that the argument of Fraser [3], p. 580 (top) showing that the interval for $\mu_{[k]}$ generated by his proof has coverage *at least* γ doesn't extend to our case, since although

$$\{R_{k-1}(x_j, j \neq l) \leq A\} => \{R_k(x_j) > A \text{ iff } x_l > A\},$$
$$\{R_{i-1}(x_j, j \neq l) \leq A\} \not=> \{R_i(x_j) > A \text{ iff } x_l > A\}.$$

Note that (if we wish to consider location parameters and not means) restriction $E_f < \infty$ in (23) can be dropped throughout this section and the results stated in terms of $\theta_{[1]}, \ldots, \theta_{[k]}$.

3. ACKNOWLEDGMENTS

Thanks are especially due to Professor Lionel Weiss for his suggestions for and guidance of Dudewicz [2], which contains these results, to Professor Robert E. Bechhofer for introducing me to ranking and selection procedures and suggesting the problem of confidence intervals for ranked means to me, and to a referee for improvement of the presentation.

APPENDIX

(A–1) THEOREM: For $i = 1, \ldots, k$ and $x \in \mathscr{R}$, $F_{\bar{X}[i]}(x) \downarrow$ as $\mu_l \uparrow$ $(l = 1, \ldots, k)$.

PROOF: Fix $l (1 \leq l \leq k)$. For $i = 1, \ldots, k$ and $x \in \mathscr{R}$,

$$\begin{aligned} F_{\bar{x}[i]}(x) &= P_\mu[\bar{X}_{[i]} \leq x] = P_\mu[\text{The } i\text{th smallest of } \bar{X}_1, \ldots, \bar{X}_k \text{ is} \leq x] \\ &= P_\mu[\text{At least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_k \text{ are} \leq x] \\ &= P_\mu[\bar{X}_l \leq x \text{ and at least } i-1 \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x] \\ &\quad + P_\mu[\bar{X}_l > x \text{ and at least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x] \\ &= P_\mu[\bar{X}_l \leq x] P_\mu[\text{At least } i-1 \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x] \\ &\quad + (1 - P_\mu[\bar{X}_l \leq x]) P_\mu[\text{At least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x] \\ &= G_n(x - \theta_l|f) P_\mu[\text{At least } i-1 \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x] \\ &\quad + [1 - G_n(x - \theta_l|f)] P_\mu[\text{At least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x]. \end{aligned}$$

Therefore, since $d\theta_l/d\mu_l = 1$,

$$\frac{d}{d\mu_l} F_{\bar{x}[i]}(x) = \frac{d}{d\theta_l} F_{\bar{x}[i]}(x) \frac{d\theta_l}{d\mu_l}$$

$$= -g_n(x-\theta_l|f)P_\mu\,[\text{At least } i-1 \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x]$$
$$+ g_n(x-\theta_l|f)P_\mu\,[\text{At least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x],$$

which is ≤ 0 iff

$$P_\mu\,[\text{At least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x]$$
$$\leq P_\mu\,[\text{At least } i-1 \text{ of } \bar{X}_1, \ldots, \bar{X}_{l-1}, \bar{X}_{l+1}, \ldots, \bar{X}_k \text{ are} \leq x].$$

LEMMA:

$$(\Lambda\text{-}2) \qquad F_{\bar{X}_{[i]}}(x) = \sum_{l=i}^{k} \frac{\binom{k}{l}}{k!} \sum_{\beta \epsilon S_k} F_{\bar{X}_{\beta(1)}}(x) \ldots F_{\bar{X}_{\beta(l)}}(x)[1-F_{\bar{X}_{\beta(l+1)}}(x)] \ldots [1-F_{\bar{X}_{\beta(k)}}(x)].$$

PROOF:

$$F_{\bar{X}_{[i]}}(x) = P\,[\text{At least } i \text{ of } \bar{X}_1, \ldots, \bar{X}_k \text{ are } \leq x]$$
$$= \sum_{l=i}^{k} P\,[\text{Exactly } l \text{ of } \bar{X}_1, \ldots, \bar{X}_k \text{ are} \leq x]$$
$$= \sum_{l=i}^{k} \frac{1}{l!\,(k-l)!} \sum_{\beta \epsilon S_k} P[\bar{X}_{\beta(1)} \leq x, \ldots, \bar{X}_{\beta(l)} \leq x \,\&\, \bar{X}_{\beta(l+1)} > x, \ldots, \bar{X}_{\beta(k)} > x]$$
$$= \sum_{l=i}^{k} \frac{\binom{k}{l}}{k!} \sum_{\beta \epsilon S_k} F_{\bar{X}_{\beta(1)}}(x) \ldots F_{\bar{X}_{\beta(l)}}(x)[1-F_{\bar{X}_{\beta(l+1)}}(x)] \ldots [1-F_{\bar{X}_{\beta(k)}}(x)].$$

REFERENCES

[1] Chambers, M. L. and C. Mack, "Confidence Limits for the Minimum of Two Normal Means; a New Inference Principle," New Journal of Statistics and Operational Research, *2*, 14–27 (1966).

[2] Dudewicz, E. J., "Estimation of Ordered Parameters," Department of Operations Research, Cornell University, Ithaca, New York, Technical Report No. 60 (Feb. 1969).

[3] Fraser, D. A. S., "Confidence Bounds for a Set of Means," Annals of Mathematical Statistics, *23*, 575–585 (1952).

[4] Gupta, S. S., "On Some Multiple Decision (Selection and Ranking) Rules," Technometrics *7*, 225–245 (1965).

[5] Lal Saxena, K. M. and Y. L. Tong, "Interval Estimation of the Largest Mean of k Normal Populations with Known Variances," J. Am. Statist. Assoc. *64*, 296–299 (1969).

* * *

A TWO-ECHELON MULTI-STATION INVENTORY MODEL FOR NAVY APPLICATIONS

S. Zacks*

The George Washington University Program in Logistics

ABSTRACT

The two inventory echelons under consideration are the depot, D, and k tender ships $E_1, \ldots, E_k$. The tender ships supply the demand for certain parts of operational boats (the customers). The statistical model assumes that the total monthly demands at the k tenders are stationary independent Poisson random variables, with unknown means $\lambda_1, \ldots, \lambda_k$. The stock levels on the tenders, at the beginning of each month, can be adjusted either by ordering more units from the depot, or by shipping back to the depot an excess stock. There is no traffic of stock between tenders which is not via the depot. The lead time from the depot to the tenders is at most 1 month. The lead time for orders of the depot from the manufacturer is L months. The loss function due to erroneous decision is comprised of linear functions of the extra monthly stocks, and linear functions of shortages at the tenders and at the depot over the N months. A Bayes sequential decision process is set up for the optimal adjustment levels and orders of the two echelons. The Dynamic Programming recursive functions are given for a planning horizon of N months.

1. INTRODUCTION

Previously, we presented some results concerning the optimal Bayes sequential inventory policy in a one-echelon and a two-echelon inventory system [2] and [3]. The one-echelon system considered in [2] represents the type of inventory model which is appropriate for the operational ship (submarine). The important feature in this model is the possible return of excess stock to the supporting base (the tender ship) after each patrol. This assumption combined with the simple cost function employed yield a relatively simple solution to the Bayes sequential decision problem.

In [3] we extended the investigation, and considered a two-echelon inventory problem. The inventory model discussed there is analogous to that of a one operational boat (the submarine) and one supporting boat (the tender). The cost structure dealt with in this study is still relatively simple. The determination of the optimal Bayes sequential policy is, however, considerably complicated. The dynamic programming method of solution, specified in that memorandum, can yield a numerical solution to the optimal decision problem, but generally requires voluminous computation. An *ad hoc* procedure has been suggested in [3], but its efficiency has not yet been investigated.

In the present study, we consider a two-echelon multi-station inventory system. The higher echelon, I, consists of a depot which obtains parts from the manufacturers. The lower echelon, II, consists of several stations (tender ships). The operational boats (submarines) assume in the present study the role of customers, which demand at certain (random) time epochs a random number of units of specified replacement parts. In this way we eliminate the need for explicit specification of the number of operational boats supported by each of the tenders, and the time schedule of their inventory replenishment. Such an information may help by incorporating it judiciously into the inventory model to further

*Presently at The University of New Mexico, Department of Mathematics and Statistics.

reduce the inventory cost. We decided, however, to treat a simpler model, realizing the difficulties and complications involved in formulating and solving a three-echelon multi-station model with a complicated time schedule of inventory replenishment. In the following section, we outline in detail the inventory model under consideration. In Section 2 we discuss the cost structure of the present inventory problem. We treat a cost function which is considerably more complicated than those of the previous studies [2], [3]. The complication in the cost structure is due to the interaction between the tenders and the depot. On the other hand, we somewhat simplify the cost function by avoiding inventory traffic directly between the stations in echelon II. This is a realistic assumption for the Navy purposes, but might be quite a restrictive assumption in other situations.

In Section 3 we set up the Bayesian decision framework for our study. This framework is almost identical to that of [3], and similar to that of [2]. In Section 4 we define recursively the functions that should be determined for the dynamic programming solution of the optimal Bayes sequential decision problem. These functions are defined and determined by the backward induction process of the dynamic programming. In principle we have thus obtained the optimal Bayes solution to the present inventory decision problem. In practice, the optimal solution requires a preparation of high-dimensional tables of functions, which are defined only implicitly. The numerical task required to prepare these tables is expected to be too difficult even with the existing fast computers. It is therefore important to develop *ad hoc* procedures which are close to the optimal one. This is the objective of our future investigations.

Finally, we should mention here the interesting study of D. Gross [1] on the inventory control of multi-location supply systems. Gross' model is very similar to ours. There are two main differences: (i) In Gross' model there is a direct inventory flow between the stations, while in our model there is none; (ii) Gross' cost structure is different. It is comprised of different cost elements, and reveals a substantially simpler form of the cost function.

2. THE INVENTORY MODEL

We consider an inventory system comprised of two echelons. Echelon I is the depot, D, and echelon II consists of k stations $E_1, \ldots, E_k$ (the tender boats). The customers are operational boats (submarines) which demand replacement parts. The operational boats receive the required parts at the stations at certain time epochs which are irrelevant for the present model. Furthermore, we do not consider in the present model the question of how many customers (submarines) are served in a given time period by a station. The only relevant statistic for the present model is the number of units of a certain part which were demanded at each of the k stations during a specified time period. Without loss of generality we specify the time period to be one month. Thus, let $X_n^{(i)}$ $(i=1,\ldots,k)$ be the number of units demanded at station i during the nth month. We denote by X_n the demand vector: $X_n=(X_n^{(i)}, \ldots, X_n^{(k)})'$. We further assume that $\{X_n^{(i)} : i=1, \ldots, k;\ n=1, 2, \ldots\}$ are independent random variables, having Poisson distributions with means λ_i $(i=1, \ldots, k)$. $0<\lambda_i<\infty$ are unknown. A complete sufficient statistic for λ_i $(i=1,\ldots,k)$ at the end of the nth month is $T_n^{(i)}=\sum_{j=1}^{n} X_j^{(i)}$. $T_n^{(1)}, \ldots, T_n^{(k)}$ are independent random variables, having Poisson distributions with means $n\lambda_i$ $(i=1, \ldots, k)$, respectively.

In the present model we assume the following inventory flow (see Figure 1).

1. At the beginning of the nth month, the depot D orders from the manufacturers Z_n units, $Z_n \geq 0$. The lead time is L months $(L \geq 1)$. That is, an order from D issued at the beginning of the nth month

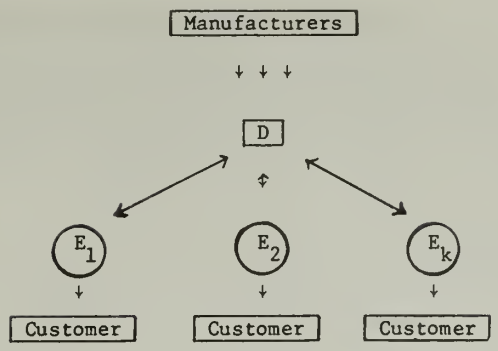


FIGURE 1

reaches the depot, D, at the *end* of the $(n+L-1)$th month. Units ordered from the manufacturers cannot be returned from the depot to the manufacturers.

2. At the beginning of the nth month station E_i of echelon II $(i=1, \ldots, k)$ orders $Y_n^{(i)}$ units from the depot D. The lead time is 1 month. The amount ordered by E_i at the beginning of the nth month arrives at E_i by the end of that month. There is no flow of material between the stations of echelon II, only through the depot. Excess of stock at the stations can be adjusted by sending units back to the depot at the beginning of a month. This is taken into account by letting $Y_n^{(i)}$ be negative. A negative value of $Y_n^{(i)}$ means a flow back from the station E_i to D.

3. There is no flow back of units from the customers to the stations. The customers demand on a one-to-one basis of failure of parts. That is, the stock levels of the customers (submarines) are specified and fixed. Demand for new parts is according to the exact number of failures of old parts.

Let Q_n designate the stock level at the depot, D, at the *end* of the nth month. Let $S_n^{(i)}$ $(i=1, \ldots, k)$ designate the stock level at station E_i at the *end* of the nth month. We designate by S_n the vector $(S_n^{(1)}, \ldots, S_n^{(k)})$. We impose on the order levels $Y_n^{(i)}$ the restriction

$$\text{(1.1)} \qquad \begin{cases} -S_{n-1}^{(i)} \leq Y_n^{(i)}, \text{ for all } i=1, \ldots, k;\ n=1, 2, \ldots \\ \sum_{i=1}^{k} Y_n^{(i)} \leq Q_{n-1}, \text{ for all } n=1, 2, \ldots \end{cases}$$

That is, each station can obviously return to D not more than available in its stock; and the total amount required by the stations from D cannot exceed the amount available at D. This constraint may create shortages at the stations due to an unexpected high demand of the customers. We assume that a shortage at the stations can always be replenished instantaneously (very short lead time) by expensive means of communication, from the depot, D, or from external resources if unavailable at the depot. This is taken into account in the cost function (discussed in the next section). There are no backlogs.

Useful relationships between the variables discussed in this model are

$$\text{(1.2)} \qquad S_n^{(i)} = (S_{n-1}^{(i)} + Y_n^{(i)} - X_n^{(i)})^+, \quad i=1, \ldots, k;\ n=1, 2, \ldots,$$

where $x^+ = \max(0, x)$, and

$$\text{(1.3)} \qquad Q_n = \left[Q_{n-1} + Z_{n-L+1} - \sum_{i=1}^{k} Y_n^{(i)} - \sum_{i=1}^{k} (S_{n-1}^{(i)} - X_n^{(i)})^- \right]^+, \quad n=1, 2, \ldots,$$

where $x^- = -\min(0, x)$, and $Z_j \equiv 0$ if $j \leq 0$. $S_0^{(i)} (i = 1, \ldots, k)$ is the initial stock level at station E_i. Q_0 is the initial stock level at D. In the sequel we consider the inventory problem for a planning horizon of N months. In this case, the depot does not order from the manufacturers after the beginning of the $(N-L)$th month. The stations do not order from the depot at the beginning of the Nth month. Thus, we have $Z_{N-j} \equiv 0$ for all $j = 0, 1, \ldots, L-1$ and $Y_N^{(i)} \equiv 0$ for all $i = 1, \ldots, k$.

3. THE COST STRUCTURE

Let $c_i (i = 1, \ldots, k)$ be the monthly carrying cost of one unit in E_i; and let p_i be the penalty cost for a shortage of one unit in E_i. Let r be the monthly carrying cost of one unit in the depot, D, and $p = \lambda r$ the penalty cost for a shortage of one unit in D. For simplicity, we assume that all costs are charged at the end of each month. The total inventory (opportunity) cost of the system at the end of the nth month is

$$
\begin{aligned}
(2.1) \qquad L_n(Q_{n-1}, S_{n-1}; X_n) = & \sum_{i=1}^{k} [c_i (S_{n-1}^{(i)} - X_n^{(i)})^+ + p_i (S_{n-1}^{(i)} - X_n^{(i)})^-] \\
& + r \left[Q_{n-1} - \sum_{i=1}^{k} Y_n^{(i)} - \sum_{i=1}^{k} (S_{n-1}^{(i)} - X_n^{(i)})^- \right]^+ \\
& + \lambda r \left[Q_{n-1} - \sum_{i=1}^{k} Y_n^{(i)} - \sum_{i=1}^{k} (S_{n-1}^{(i)} - X_n^{(i)})^- \right]^- , \quad n = 1, \ldots, N.
\end{aligned}
$$

The objective is to determine the order levels $Y_n = (Y_n^{(1)}, \ldots, Y_n^{(k)})$ and Z_n at the beginning of each month, so as to minimize the expected total cost:

$$
(2.2) \qquad R = E \left\{ \sum_{n=1}^{N} L_n (Q_{n-1}, S_{n-1}; X_n) \right\}.
$$

The expectation (2.2) depends on the unknown Poisson parameters $(\lambda_1, \ldots, \lambda_k)$. In the Bayesian decision framework, which we adopt, the decisions depend initially on the particular prior assumptions concerning $(\lambda_1, \ldots, \lambda_k)$; however, as information about the actual demand is accumulated, the dependence of the decisions on the prior assumptions weakens.

4. THE BAYESIAN FRAMEWORK

In the Bayesian framework we postulate a certain prior distribution for $(\lambda_1, \ldots, \lambda_k)$, which we convert sequentially to a posterior distribution, given T_n $(n = 1, 2 \ldots)$. We also need the marginal posterior distribution of X_{n+1} given T_n. This marginal posterior distribution is the mixture of the Poisson distributions of X_{n+1} given $(\lambda_1, \ldots, \lambda_k)$, with respect to the posterior distribution of $(\lambda_1, \ldots, \lambda_k)$ given $T_n = (T_n^{(1)}, \ldots, T_n^{(k)})$. As in the previous studies [2] and [3], we let $(\lambda_1, \ldots, \lambda_k)$ be priorly independent, having the same gamma distribution $G(\tau^{-1}, \nu)$. Accordingly $\{X_{n+1}^{(i)}: i = 1, \ldots, k\}$ are posteriorly independent, i.e., the marginal joint posterior density is the product of the individual marginal posterior densities.

Let N.B. (ν, ψ) designate a negative binomial distribution, with density function:

$$
(3.1) \qquad f(x \mid \nu, \psi) = (1-\psi)^{\nu} \binom{\nu + x - 1}{x} \psi^x, \; x = 0, 1, \ldots \; 0 < \psi < 1 \text{ and } 0 < \nu < \infty.
$$

It is easy to verify (see [2]) that under the above assumption, the marginal posterior distribution of $X_{n+1}^{(i)}$ given $T_n^{(i)}$ $(i = 1, \ldots, k)$ is the N.B. $(\nu + T_n^{(i)}, \psi_n)$, where $\psi_n = \tau/(1 + (n+1)\tau)$. In the sequel we shall apply the notation $E_{x \mid \nu_n^{(i)}, \psi_n} \{\phi(X)\}$ to designate the expectation of the function $\phi(X)$ with respect to the N.B. $(\nu_n^{(i)}, \psi_n)$ distribution; where $\nu_n^{(i)} = \nu + T_n^{(i)}$ $(n = 1, 2, \ldots; i = 1, \ldots, k)$.

5. THE DYNAMIC PROGRAMMING SOLUTION

In the present section we prescribe the backward induction process of dynamic programming. After the nth month we are given the statistic T_n, and the stock information $\zeta_n = (Q_n, S_n)$, and in case of $L \geq 2$, the pending orders $\xi_{n,L} = (Z_{n-L+2}, \ldots, Z_n)$. In case of $L=1$ there are no pending orders at the end of each month. As we shall see in the sequel, the Bayes optimal decision concerning (Z_{n+1}, Y_{n+1}), at the beginning of the $(n+1)$st month, is a function of $(T_n, \zeta_n, \xi_{n,L})$.

Starting with the last month, we define the expected posterior risk for the Nth month as

$$\rho_1(Q_{N-1}, S_{N-1}^{(1)}, \ldots S_{N-1}^{(k)}; \nu_{N-1}^{(1)}, \ldots, \nu_{N-1}^{(k)}) = E_{X_N|\nu_{N-1},\psi_{N-1}}\{L_N(\zeta_{N-1}, X_N)\}$$

$$= \sum_{i=1}^{k} E_{X_N^{(i)}|\nu_{N-1}^{(i)},\psi_{N-1}}\{c_i(S^{(i)}{}_{N-1} - X_N^{(i)})^+ + p_i(S_{N-1}^{(i)} - X_{N-1}^{(i)})^-\}$$

$$+ rE_{X_N|\nu_{N-1},\psi_{N-1}}\{[Q_{N-1} - \sum_{i=1}^{k}(S_{N-1}^{(i)} - X_N^{(i)})^-]^+ + \lambda[Q_{N-1} - \sum_{i=1}^{k}(S_{N-1}^{(i)} - X_N^{(i)})^-]^-\}. \tag{4.1}$$

We notice that no orders are issued at the beginning of the Nth month, and therefore $\rho_1(Q_{N-1}, S_{N-1}; \nu_{N-1})^*$ given by (4.1) is a minimal Bayes function.

By the process of backward induction we determine for the $(N-1)$st month the function

$$R_2(Q_{N-2}, S_{N-2}; \nu_{N-2}, y_1, \ldots, y_k, z_1)$$

$$= E_{X_{N-1}|\nu_{N-2}}\{L_{N-1}(Q_{N-2}, S_{N-2}; X_{N-1}, y_1, \ldots, y_k)$$

$$+ \rho_1([Q_{N-2} + z_1 - \sum_{i=1}^{k} y_i - \sum_{i=1}^{k}(S_{N-2}^{(i)} - X_{N-1}^{(i)})^-]^+, (S_{N-2}^{(1)} + y_1 - X_{N-1}^{(1)})^+, \ldots,$$

$$(S_{N-2}^{(k)} + y_k - X_{N-1}^{(k)})^+; \nu_{N-2} + X_{N-1})\}. \tag{4.2}$$

This function yields the expected posterior cost of the $(N-1)$st plus that of the Nth month, given $(Q_{N-2}, S_{N-2}, T_{N-2})$, when $y_1, \ldots, y_k$ are the order levels of the stations $E_1, \ldots, E_k$ at the beginning of the $(N-1)$st month; and z_1 is the order level of the depot D at the beginning of the $(N-L)$th month. We then determine the value of $(y_1, \ldots, y_k)$ which minimizes (4.2) under the constraint (1.1). This minimizing vector $(y_1^0, \ldots, y_k^0)$ represents the optimal order level of the stations, at the beginning of the $(N-1)$st month. It is a function of $(\zeta_{N-2}, \nu_{N-2}, z_1)$. The minimal Bayes function corresponding to (4.2) is:

$$\rho_2(Q_{N-2}, S_{N-2}; \nu_{N-2}, z_1) = \inf_{(y_1, \ldots, y_k)\in\phi_{N-2}} R_2\{Q_{N-2}, S_{N-2}; \nu_{N-2}, y_1, \ldots, y_k, z_1\}, \tag{4.3}$$

where $\phi_{N-2} = \left\{(y_1, \ldots, y_k) : y_i \geq -S_{N-2}^{(i)} \text{ for all } i=1, \ldots, k, \text{ and } \sum_{i=1}^{k} y_i \leq Q_{N-2}\right\}$. The value of $Z_{N-1} = z_1$ cannot be fixed now unless $L=1$.

Generally, for every $j=1, \ldots, L$ we define

$$R_{j+1}(Q_{N-j-1}, S_{N-j-1}; \nu_{N-j-1}, y_1, \ldots, y_k, z_1, \ldots, z_j) = E_{X_{N-j}|\nu_{N-j-1}}$$

$$\left\{L_{N-j}(Q_{N-j-1}, S_{N-j-1}; X_{N-j}, y) + \rho_j\left(\left[Q_{N-j-1} + z_j - \sum_{i=1}^{k} y_i - \sum_{i=1}^{k}(S_{N-j-1}^{(i)} - X_{N-j}^{(i)})^-\right]^+,\right.\right.$$

$$\left.\left.(S_{N-j-1} + y - X_{N-j})^+; \nu_{N-j-1} + X_{N-j}, z_1, z_2, \ldots, z_{j-1}\right)\right\}. \tag{4.4}$$

*Symbols without superscripts designate k-dimensional vectors.

where the vector $(S_{N-j-1}+y-X_{N-j})^+$ has the components $(S^{(i)}_{N-j-1}+y_i-X^{(i)}_{N-j})^+$, $i=1, \ldots, k$. After the function (4.4) has been determined, we find the vector $(y_1^0, \ldots, y_k^0)$ which minimizes (4.4) under the constraint (1.1). This is the optimal order level of the stations $E_1, \ldots, E_k$ at the beginning of $(N-j)$th month. We notice that it depends not only on the current stock level (Q_{N-j-1}, S_{N-j-1}) but also all on the pending orders. As mentioned earlier, z_1, is the amount ordered at the beginning of the $(N-L)$th month. Similarly, $z_2=Z_{N-L-1}$, $z_3=Z_{N-L-2}$, and so on. Next step is to determine the minimal Bayes function, which is

$$\rho_{j+1}(Q_{N-j-1}, S_{N-j-1}; \nu_{N-j-1}, z_1, \ldots, z_j) = \inf_{(y_1, \ldots, y_k)\in\phi_{N-j-1}} R_{j+1}(Q_{N-j-1}, S_{N-j-1}; \nu_{N-j-1}, y_1, \ldots, y_k, z_1, \ldots, z_j), \tag{4.5}$$

where

$$\phi_{N-j-1}=\left\{(y_1, \ldots, y_k) : y_i \geq -S^{(i)}_{N-j-1} (i=1, \ldots, k) \text{ and } \sum_{i=1}^{k} y_i \leq Q_{N-j-1}\right\}. \tag{4.6}$$

For $j=L$, we define also

$$\rho^*_{L+1}(Q_{N-L-1}, S_{N-L-1}; \nu_{N-L-1}, z_2, \ldots, z_L) = \inf_{0\leq z_1<\infty} \rho_{L+1}(Q_{N-L-1}, S_{N-L-1}; \nu_{N-L-1}, z_1, \ldots, z_L). \tag{4.7}$$

The value of z_1, for which the infinum in (4.7) is attained, is the optimal order level of the depot, D, at the beginning of the $(N-L)$th month.

Finally, for every $a=1, \ldots, N-(L+1)$, we define

$$\begin{aligned} R_{L+1+a}(Q_{N-(L+1+a)}, S_{N-(L+1+a)}; \nu_{N-(L+1+a)}, y_1, \ldots, y_k, z_{1+a}, \ldots, z_{L+a}) \\ = E_{X_{N-(L+a)}|\nu_{N-(L+1+a)}}\{L_{N-(L+a)}(Q_{N-(L+1+a)}, S_{N-(L+1+a)}; X_{N-(L+a)}, y) \\ + \rho^*_{L+a}([Q_{N-(L+1+a)}+z_{L+a}-\sum_{i=1}^{k} y_i-\sum_{i=1}^{k} (S^{(i)}_{N-(L+1+a)}-X^{(i)}_{N-(L+a)})^-]^+, \\ (S_{N-(L+1+a)}+y-X_{N-(L+a)})^+; \nu_{N-(L+1+a)}+X_{N-(L+a)}, z_{1+a}, \ldots, z_{L+a-1})\}. \end{aligned} \tag{4.8}$$

We determine the values of y_i which minimize (4.8), and the optimal z_{1+a}. Thus, the corresponding minimal Bayes function is

$$\begin{aligned} \rho^*_{L+1+a}(Q_{N-(L+1+a)}, S_{N-(L+1+a)}; \nu_{N-(L+1+a)}, z_{2+a}, \ldots, z_{L+a}) \\ = \inf_{0\leq z_{1+a}<\infty} \inf_{(y_1, \ldots, y_k)\in\phi_{N-(L+1+a)}} R_{L+1+a}(Q_{N-(L+1+a)}, S_{N-(L+1+a)}; \\ \nu_{N-(L+1+a)}, y_1, \ldots, y_k, z_{1+a}, \ldots, z_{L+a}). \end{aligned} \tag{4.9}$$

The set $\phi_{N-(L+1+a)}$ is defined like in (4.6), in which j is replaced by $L+a$. As before, the values of $y_1, \ldots, y_k$ in $\phi_{N-(L+1+a)}$ which minimize (4.8) are the optimal order levels of $E_1, \ldots, E_k$ at the beginning of the $N-(L+a)$th month. The value of z_{1+a} for which (4.9) is attained, is the optimal order level of the depot, D, at the beginning of the $N-(L+a)$th month. All these order levels depend on the pending orders of the depot $(z_{2+a}, \ldots, z_{L+a})$.

By this process we determine recursively the minimal Bayes functions $\rho_1, \rho_2, \ldots, \rho_L, \rho^*_{L+1}, \ldots, \rho^*_N$. The function ρ^*_N depends on the initial stock level at D, Q_0; the initial stock levels at the stations $S^{(i)}_0 (i=1, \ldots, k)$, and the prior assumptions concerning ν and $\tau \equiv \psi_0$. For given values of ν and τ we can determine the initial stock levels Q_0 and S_0 to minimize $\rho^*_N(Q_0, S_0; \nu)$.

The minimal Bayes functions given by (4.11) and (4.12) are quite complicated. A numerical solution requires the tabulation of these N functions for all possible realizations of the $(2k+1)$ dimensional vector $(Q_j, S^{(1)}_j, \ldots, S^{(k)}, \nu^{(1)}_j, \ldots, \nu^{(k)}_j)$ $(j=0, \ldots, N)$ and the possible values of the pending orders. Thus, even for $k=2$ the Dynamic Programming procedure as formulated here might prove to be too complicated for practical purposes. In the case of one station in echelon II, i.e., $k=1$, and 1 month lead time from the manufacturer to the depot, $L=1$, all these functions are simplified substantially. A numerical solution in this particular case is generally attainable with no considerable difficulty.

REFERENCES

[1] Gross, D., "Generalized Inventory Control in Multilocation supply Systems," *Multistage Inventory Models and Techniques* (Editors—H. E. Scarf, D. M. Gilford, and M. W. Shelly) (Stanford University Press, Stanford, California, 1963).

[2] Zacks, S., "Bayes Sequential Design of Stock Levels," Naval Research Logistics Quarterly *16*, 143–155 (June 1969).

[3] Zacks, S., "Bayes Sequential Design of Stock Levels in a Two-Echelon Inventory System," Logistics Research Project, The George Washington University. Technical Memorandum Serial TM–22016. (1968).

* * *

A GAME THEORY APPROACH TO LOGISTICS ALLOCATION*

Sidney Moglewer and Carl Payne†

McDonnell Douglas Corporation
Douglas Aircraft Company
Long Beach, California

ABSTRACT

A simple model containing elements common to all logistics pipeline allocations is presented. This model is based on a two-sided game.

1. INTRODUCTION

The McNamara era in the Department of Defense was characterized by a major buildup of airlift and sealift forces. In quantitative terms, the planned program would provide for an elevenfold increase of strategic airlift in FY 1973 over 1961 [5]. The Fast Deployment Logistics Ship concept and land prepositioning concepts also evolved during this era. Underlying the resource allocation to these programs has been the notion that rapid deployment of U.S. forces was always advantageous. For instance, Paul R. Ignatius, then Assistant Secretary of Defense, at an NSIA Advanced Planning Briefing Conference on 3 March 1965 in Los Angeles, stated: "Closely related to the size and composition of our general purpose forces is the capability of moving them promptly to wherever they may be needed. Obviously, the more rapidly we can introduce substantial U.S. ground and air forces into an area under attack, the smaller the total U.S. force commitment is likely to be." It is the thesis of this paper that this is not so obvious, and in fact, in many instances may be questionable. Although there have been many recent past studies in the area of airlift and sealift, it is felt that many of these studies do not adequately take into account the response of an intelligent enemy. For instance, a linear programming model to optimize a least-cost transport fleet mix requires as an input a specified force deployment requirement in a given period of time [7]. Inasmuch as a deployment is very sensitive to the response of the enemy opposing this deployment, the critical element for the decision-maker is left out of the model. Furthermore, certain simulations to evaluate the balance between speed and cost in deployment situations have *a priori* assumptions built into them on the value of rapid deployment [2]. None of the studies to date have delineated the nature of the decision-making structure concerned with the value of rapid deployment. Consequently, it is felt that the strategic and tactical lift problem has not yet been "studied to death." In fact, critical decision-making elements are still unresolved.

This paper presents a simple model containing elements common to all military logistics pipeline allocations. This model utilizes the framework of a two-sided game, one wherein the enemy is free to do whatever is best for him; and, at any time can take advantage of his opponent's mistakes. The

*This paper was presented at the 21st Military Operations Research Symposium, Colorado Springs, Colorado, June 11–13, 1968.

§Now at Calif. State College, Long Beach, Calif. Dept. of Operations Research and Statistics.

results of this model provide insight into the decision-making structure of the deployment problem and suggestions for further areas of study. Like any model, including the most complex simulation or war game, this model is only an abstraction of the real world. Thus, what is important is not the specific numerical results, but rather the insight obtained by the model into a qualitative description of the structure of the logistics allocation decision. This can provide the basis for more detailed analysis of airlift and sealift resource allocation and help mold the intuitive assessment of the real world problem.

2. THE LOGISTICS ALLOCATION DECISION

Logistics pipeline allocation decisions are a sensitive tradeoff between cargo destined for combat and that destined to further build up the pipeline. There is a natural division of cargo into three distinct categories:

1. *Deployment* (The combat forces consisting of personnel, guns, tanks, vehicles, and the like) These are usually low density, high value, mobile items from a logistics point of view.

2. *Resupply* (Materials needed to support deployed forces) Major items are petroleum, ammunition, spare parts, and subsistence items. These are usually medium-density, low-value items that are packaged in a standard form (pallets, containers, etc.) and require materials handling equipment.

3. *Facilities Support* (Materials used to build up the pipeline) Included in this category are construction equipment and vehicles for building ports, runways, warehouses, and the like. It also includes support for logistics and service commands. These items may be high-density, low-value items and oftentimes require special handling.

Due to the above physical differences combined with variations in demand, transportation needs can vary widely. When the competitive effects of a hostile enemy are also considered, the complexity of the decision to select an optimum fleet of transport vehicles becomes apparent. The possibilities open to the enemy may require the addition of cargo for pipeline defense and cargo to interdict the enemy pipeline. These additions can vary widely from sophisticated missile batteries to sandbags. Furthermore, vulnerability factors for the transportation vehicles and their facilities may become significant. For transport fleet mix optimization, the mission-oriented cost effectiveness study is presently severely limited by the lack of a methodological approach for selection of suitable evaluation missions. The real world mission is not a clearly defined deployment or resupply mission, but rather a mix of deployment and resupply and other things, as demanded by the contingency situation.

The problem faced by the military commander, constrained by a finite logistics pipeline, at any given instant of time is how to allocate the different classes of cargo through the pipeline. He would like to allocate a large amount of deployment cargo in order to augment his combat forces, not knowing what deployment forces the enemy is committing. At the same time, it is necessary to allocate sufficient resupply cargo in order to support present forces and for stockpiling purposes. Furthermore, he would like to allocate a large amount of facilities support cargo in order to build a larger pipeline in the future. In addition, it may be necessary to allocate cargo for pipeline defense and for interdiction of the enemy pipeline. All these allocation decisions must be made in ignorance of the enemy's decision. The most cost effective transportation fleet mix is very sensitive to the nature of these logistics allocation decisions by both sides.

3. THE LOGISTICS ALLOCATION MODEL

A two variable differential game model has been initially developed to address this problem. The model considers the allocation of pipeline capacity between two factors, deployment and facilities

support. Resupply is assumed to be a function of total surviving deployed forces. (Stockpiling is not incorporated into this study.)

Each player's payoff is a function of his deployment allocation. There are many forms that a specific payoff function can take within the overall goal of winning the war and still be representative of a particular real world situation. A suitable objective might be to destroy the enemy forces as rapidly as possible. Another might be to minimize losses. A third might be to maximize battlefield force superiority prior to negotiations. In fact, during a given campaign there may be a sequence of different payoffs with payoff asymmetry between opponents. What is important is not so much the specific payoff chosen, but rather those elements of the solution insensitive to most payoffs. This will determine the basic decision structure of the problem.

Although it is planned to investigate this aspect of the problem at a later date, for the purposes of this paper, a battlefield superiority type of payoff will be chosen. Each player's objective is to maximize his deployment superiority. Battlefield attrition of deployed forces is assumed to follow a Lanchester square law [4]. It has been found that this law holds for some individual battles [1], although evidence to the contrary also exists [6]. In this model, the Lanchester square law is essentially nothing more than a logistics consumption function, and if another more accurate function is discovered, it can be substituted into the analysis. In this initial approach, pipeline defense and interdiction of the pipeline by either side will not be considered.

The model assumes a battle zone to which two opponents bring deployment forces and resupply via logistics pipelines. The termini of the pipelines are assumed to be in protected areas. The deployed forces engage in combat on the battlefield, each side sustaining losses at a rate proportional to the opponent's deployment.

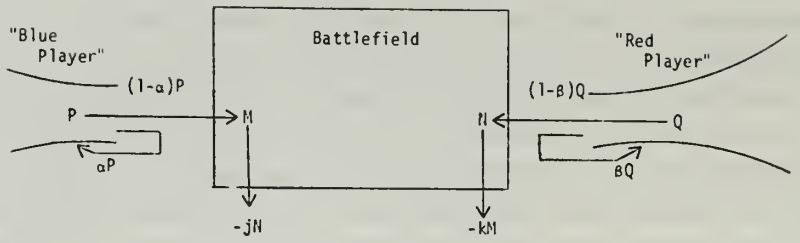
Facilities support brought through the pipeline is assumed to increase the size of the pipeline. This is representative of the situation in which lack of port facilities, docks, warehouses, materials handling equipment, airstrips, and the like along the pipeline restrict its size, and these items must be brought through the pipeline in order to increase its capacity. A large source of supplies input to the pipeline is assumed. Figure 1 depicts the situation.

The increase in pipeline capacity is assumed to be proportional to the amount of equipment and supplies invested in it. The proportionality constants, which are measures of the productivity of capital, are g and h, respectively, for Blue and Red. Although productivity should tend to follow a law of diminishing return, and the productivity coefficients may be different in the short run as compared to a longer time span, for purposes of this first investigation a constant linear relationship will be used.

The following differential equations describe the progress of the battle:

$$\dot{M} = (1-\alpha)P - jN.$$
$$\dot{N} = (1-\beta)Q - kM.$$
$$\dot{P} = g\alpha P.$$
$$\dot{Q} = h\beta Q.$$

The objective of each side is assumed to be the maximization of its deployment superiority. Payoff $= (M-N)$ at the termination of conflict. Blue is the maximizing player and Red the minimizing player. Each player attempts to achieve his objective by choosing the value of α or β as a function of M, N, P, and Q at any time. Either player can take advantage of poor strategies by the other player. Both players know the values of M, N, P, and Q at any time, but not the value of α or β being chosen by the other at that time. Both players have a complete history of past play.



Explanation of Symbols

M = Blue's deployed force on the battlefield at any time.

N = Red's deployed force on the battlefield at any time.

P = Blue's pipeline capacity at any time.

Q = Red's pipeline capacity at any time.

α = The proportion of his pipeline at any time that Blue allocates to facilities support.

β = The proportion of his pipeline at any time that Red allocates to facilities support.

j = The amount of Blue's deployment destroyed per unit time per unit of Red's deployed force.

k = The amount of Red's deployment destroyed per unit time per unit of Blue's deployed force.

FIGURE 1. Logistics allocation model

Termination of the game is taken as the reduction of a player's deployment to zero. Without loss of generality, this player can be labeled Red. Then the termination is given by $N_0=0$, and the payoff reduces to M_0 the size of Blue's deployment surviving at the end of the game. The above assumption is primarily for mathematical convenience and other stopping rules could be applied without loss of generality. For instance, it has been conjectured that when a conventional force suffers an attrition of 15–20 percent of its deployment, it becomes too disorganized to be effective militarily.

Solution of the Game:

The Kinematic equations are given as follows:

$$f_M=\dot{M}=(1-\alpha)P-jN.$$
$$f_N=\dot{N}=(1-\beta)Q-kM.$$
$$f_P=\dot{P}=g\alpha P.$$
$$f_Q=\dot{Q}=h\beta Q.$$

Note that all variables and constants are positive, and additionally

$$0\leq\alpha\leq 1,\ 0\leq\beta\leq 1.$$

By letting V be the Value of the game (payoff under optimal play by both sides), and $V_M=\partial V/\partial M$, etc., the set of path equations is completed as follows:

$$\dot{V}_M = -V_M \partial f_M/\partial M - V_N \partial f_N/\partial M - V_P \partial f_P/\partial M - V_Q \partial f_Q/\partial M = kV_N.$$
$$\dot{V}_N = jV_M.$$
$$\dot{V}_P = -V_M(1-\bar{\alpha}) - V_P g\bar{\alpha}.$$
$$\dot{V}_Q = -V_N(1-\bar{\beta}) - V_Q h\bar{\beta}.$$

The barred variable indicates optimal play. Note that the payoff, being terminal, does not appear in the path equations.

The Retrogressive Path Equations are then:

$$\dot{V}_M = -kV_N$$
$$\dot{V}_N = -jV_M$$
$$\dot{V}_P = V_M(1-\bar{\alpha}) + V_P g\bar{\alpha}$$
$$\dot{V}_Q = V_N(1-\bar{\beta}) + V_Q h\bar{\beta}$$
$$\dot{M} = jN - (1-\bar{\alpha})P$$
$$\dot{N} = kM - (1-\bar{\beta})Q$$
$$\dot{P} = -g\bar{\alpha}P$$
$$\dot{Q} = -h\bar{\beta}Q$$

The Main Equation, Eq. (1), is

$$\min_{\beta} \max_{\alpha} [(V_P g - V_M)P\alpha + (V_Q h - V_N)Q\beta + V_M P - V_M jN + V_N Q - V_N kM] = 0, \tag{1}$$

from which the switching functions for $\bar{\alpha}$ and $\bar{\beta}$ are determined.

$$\bar{\alpha} = \begin{cases} 1 \text{ if } V_P g - V_M > 0 \\ 0 \text{ if } V_P g - V_M < 0 \end{cases}$$

and

$$\bar{\beta} = \begin{cases} 1 \text{ if } V_Q h - V_N < 0 \\ 0 \text{ if } V_Q h - V_N > 0. \end{cases}$$

Refer to Isaacs [3] for a more complete description of the approach of differential games.

The terminal surface is represented parametrically by

$$M = r,$$
$$P = s,$$
$$Q = w,$$

and

$$N = 0 \text{ (zero)}.$$

Payoff: $H = M = r$

$$\frac{\partial H}{\partial r} = 1 \qquad \frac{\partial H}{\partial s} = 0 \qquad \frac{\partial H}{\partial w} = 0,$$

also

$$\frac{\partial H}{\partial r}=\frac{\partial V}{\partial M}\frac{\partial M}{\partial r}+\frac{\partial V}{\partial N}\frac{\partial N}{\partial r}+\frac{\partial V}{\partial P}\frac{\partial V}{\partial r}+\frac{\partial V}{\partial Q}\frac{\partial Q}{\partial r},$$

and similarly for

$$\frac{\partial H}{\partial s},\frac{\partial H}{\partial w}.$$

Taking partial derivatives of the parametric equations of the terminal surface and making substitutions, we obtain

$$V_{Mo}=1 \qquad V_{Po}=0 \qquad V_{Qo}=0,$$

where V_{Mo} is V_M on the terminal surface, etc.

The Main Equation, Eq. (2), is

$$V_{Mo}[(1-\bar{\alpha})s]+V_{No}[(1-\bar{\beta})w-kr]+V_{Po}g\bar{\alpha}s+V_{Qo}h\bar{\beta}w=0; \tag{2}$$

by substituting V_{Mo}, V_{Po}, and V_{Qo}

$$(1-\bar{\alpha})s+V_{No}[(1-\bar{\beta})w-kr]=0$$

converting back to original state variables,

$$V_{No}=\frac{(1-\bar{\alpha})P_o}{kMo-(1-\bar{\beta})Qo}.$$

The switching function for ā on the terminal surface is $V_{Po}g-V_{Mo}=-1$, and therefore $\bar{\alpha}=0$.

The switching function for $\bar{\beta}$ on the terminal surface is $V_{Qo}h-V_{No}=-V_{No}$.

From Eq. (2)

$$P_o+V_{No}[(1-\bar{\beta})Q_o-kMo)]=0,$$

but

$$(1-\bar{\beta})Qo-kMo=\dot{N} \text{ (2nd kinematic equation).}$$

Since the terminal surface is $N=0$, $\dot{N}$ must be negative just prior to the end of the game. For $ME-2$ to hold, V_{No} must be positive. The switching function for $\bar{\beta}$ is then negative, and $\bar{\beta}=1$.

The result appears strange at first. Why should Red conclude the game by allocating all of his pipeline capacity to increasing the size of the pipeline? Considering the Kinematic equations, when N is small, M increases approximately linearly with time if Blue is allocating all of his capacity to deployment (which he does). Red cannot reduce this build-up by attrition of Blue's forces, because Red's force is too small. Under these circumstances, Red can best achieve his objective of minimizing M by minimizing the length of time that Blue has to build his forces. In effect, Red surrenders. But since the model does not provide for this action, the closest possible allowed action is to *not deploy*.

Integrating the Retrogressive Path Equations and Employing the above boundary conditions yields equations for the state variables (M, N, P, and Q) and the derivatives of the Value (V_M, V_N, V_P, and V_Q), as functions of time. These are then substituted into the control variable switching functions to find the time at which each player should change his control variable.

At this time new boundary conditions are found, and the new *RPE* integrated in the next region. The details of this process are quite lengthy and involved. Results are summarized in Table 1.

4. THE DECISION-MAKING STRUCTURE

The interpretation of the results of Table 1 is quite interesting. Figures 2 and 3 present the optimal strategies of Blue and Red in retrograde time (i.e., time proceeds from right to left on the horizontal axis and the end of the game is at the origin.) The sensitivity of logistics pipeline allocation decisions to the relationship of the pipeline productivity coefficient and the battlefield attrition rate is clearly shown. Furthermore, optimal strategies are generally at extreme points, either all pipeline buildup or

TABLE 1. *Summary of Control Variable Changes*

Blue			
Case	Region	Time interval	Control variable
$\sqrt{jk} \leq g$	ALL	$0 \leq t < t_1$ $t_1 < t$	$\bar{\alpha}=0$ $\bar{\alpha}=1$
$\sqrt{jk} > g$	$Po/Mo \leq \sqrt{jk-g^2}$	$0 \leq t$	$\bar{\alpha}=0$
	$\sqrt{jk-g^2} < \frac{Po}{Mo} < \sqrt{jk}$	$0 \leq t < t_1$ $t_1 < t < t_2$ $t_2 < t$	$\bar{\alpha}=0$ $\bar{\alpha}=1$ $0 < \bar{\alpha} < 1$
	$Po/Mo \geq \sqrt{jk}$	$0 \leq t < t_1$ $t_1 < t$	$\bar{\alpha}=0$ $\bar{\alpha}=1$

where

$$t_1 = \frac{1}{\sqrt{jk}} \ln \left[\frac{gPo/Mo - \sqrt{jk(g^2 + Po^2/Mo^2 - jk)}}{(\sqrt{jk}-g)(\sqrt{jk}-Po/Mo)} \right],$$

which reduces to

$$t_1 = \frac{1}{\sqrt{jk}} \ln \left[1 + \frac{\sqrt{jk}}{Po/Mo} \right] \text{ when } g = \sqrt{jk},$$

and

$$t_1 = \frac{1}{\sqrt{jk}} \ln \left[1 + \frac{\sqrt{jk}}{g} \right] \text{ when } Po/Mo = \sqrt{jk},$$

t_2 must be found numerically.

TABLE 1. *Summary of Control Variable Changes*—Continued

Red			
Case	Region	Time interval	Control variable
$h \leq \sqrt{jk}$	$Po/Mo \geq \sqrt{jk}$	$0 \leq t$	$\bar{\beta}=1$
	$Po/Mo < \sqrt{jk}$	$0 \leq t < t_1'$ $t_1' < t$	$\bar{\beta}=1$ $\bar{\beta}=0$
$h > \sqrt{jk}$	$Po/Mo \geq \sqrt{jk}$	$0 \leq t$	$\bar{\beta}=1$
	$Po/Mo < \sqrt{jk}$	$0 \leq t < t_1'$ $t_1' < t < t_2'$ $t_2' < t$	$\bar{\beta}=1$ $\bar{\beta}=0$ $\bar{\beta}=1$

where

$$t_1' = \frac{1}{\sqrt{jk}} \ln \left[\frac{\sqrt{jk} + Po/Mo}{\sqrt{jk} - Po/Mo} \right]^{1/2}$$

and

$$t_2' = \frac{1}{\sqrt{jk}} \ln \left[\frac{h + \sqrt{jk}}{h - \sqrt{jk}} \right] \left[\frac{\sqrt{jk} + Po/Mo}{\sqrt{jk} - Po/Mo} \right]^{1/2}$$

all deployment (except for a narrow region for Blue where he may pursue a mix of both pipeline and deployment allocations). These can be referred to as "Bang-bang" strategies and are primarily due to the linearity of the kinematic equations in the decision variables. The solution clearly indicates that under certain conditions it does not pay to introduce deployed forces too rapidly; rather, it is better to invest in the future by first building up the pipeline prior to a later battlefield deployment.

For purposes of discussion, let us call the strategy of first building up the pipeline and then deploying a massive deployment strategy. When deployment occurs at the beginning of the game, call this a rapid deployment strategy. It should be realized that in the complex real world situation, even with a massive deployment strategy, there may be very real requirements for the rapid deployment of forces. But the objectives of these forces would be for pipeline defense and other items not incorporated into this present model. Even so, the model clearly questions the notion that rapid introduction of forces into a battle area always implies a smaller total force commitment. When productivity is high relative to attrition, a massive deployment strategy should be pursued. As productivity becomes lower, rapid deployment policies should become optimal.

There is further interesting insight to be gained from an examination of the optimal strategy spaces for both players. Any given scenario represents only one point in the continuum of points represented by Figs. 2 and 3; however, all scenarios for the Blue player can be classified as belonging to the rapid-deployment, massive-deployment, or the mixed-allocation region. Consequently, there are three basic

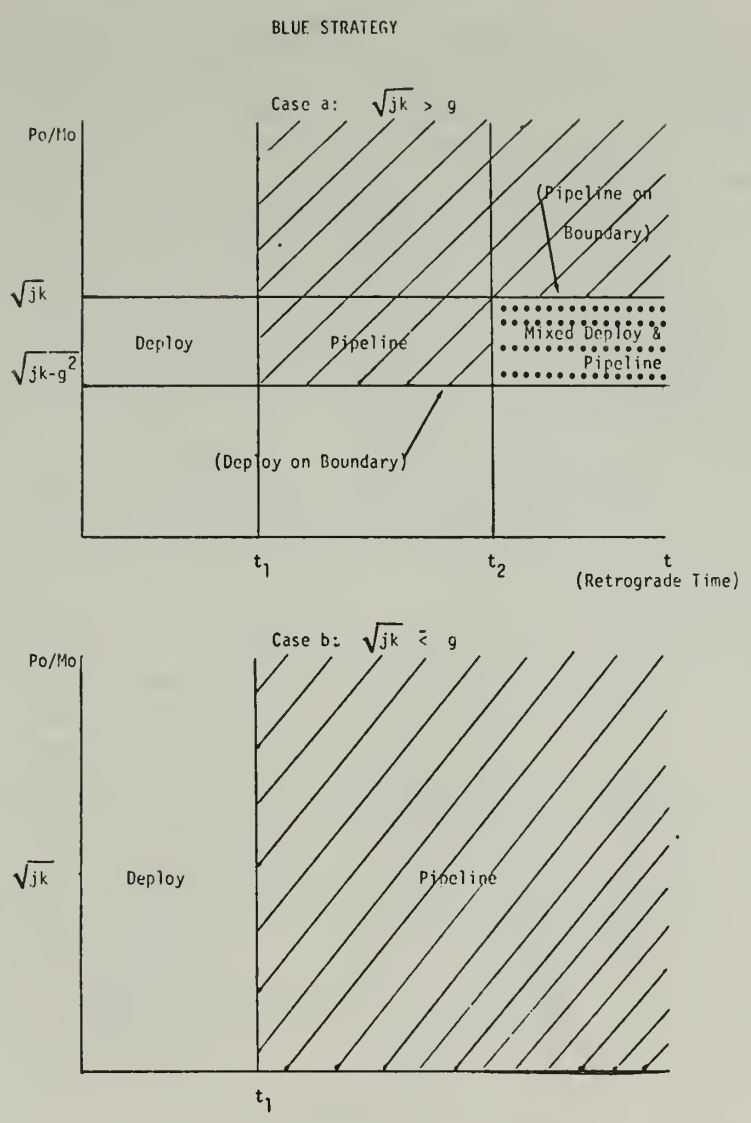


FIGURE 2. Blue strategy

force structure optimizations for the Blue player, one for each region. If Blue does not have the resources for an optimal force structure for each region, he should resort to a decision theory approach which generally implies optimizing for one region with acceptable risk in the other regions. Since this technique has apparently not been used for present airlift and sealift force structure decisions, these decisions can be questioned from the point-of-view of game theory.

Although optimal choices are few in number, the commander's operational decision is difficult to determine. For any given scenario estimate (Lanchester attrition parameter j and k and logistics productivity coefficients g and h) the commander should, in general, select either a massive or a rapid deployment based upon initial estimates of both sides' deployed forces M and N and both sides' pipeline sizes P and Q. The structure of this decision, however, even for this simple model, is extremely complex and has not yet been solved. The significane of this is that the game theory model may provide great benefits in resource allocation and force structure decisions, but with present state-of-the-art it is not likely to provide quantitative guidance for the commander's tactical decisions. By use of the

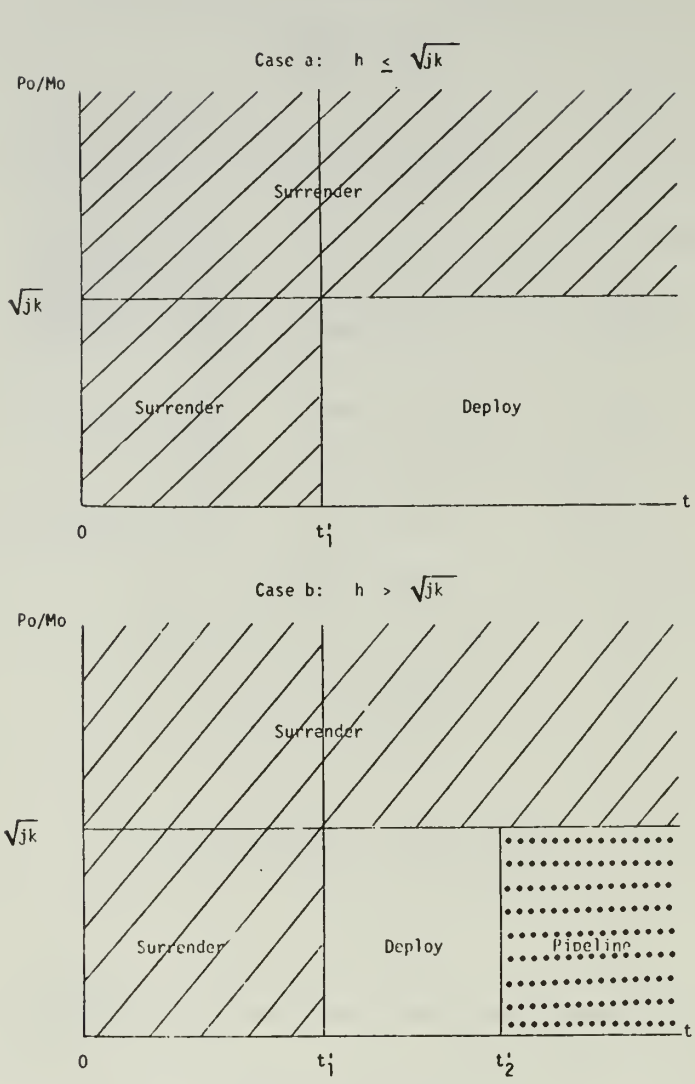


FIGURE 3. Red strategy

game theory approach, however, force structure decisions can be separated from tactics in the field. As a further complication to the commander's decision, there is at present no effective means of a priori estimation of the Lanchester attrition coefficients. Furthermore, for any given scenario, merely by varying the intensity of the battle (Lanchester coefficients j and k), it is possible to go from one strategy region to another, all other things being constant. Therefore, it does appear that military genius will still prevail on the battlefield—if not unduly handicapped by poor force structure decisions.

5. IMPLICATIONS OF THE MODEL

The basic implication of the model is the emphasis upon either massive or rapid deployment, with profound meaning for force structure decisions. Rapid deployment criteria are generally associated with a small enemy pipeline and low enemy pipeline productivity. These factors are more important than the amount of initially deployed enemy forces prior to the engagement. Furthermore, the model relates pipeline buildup decisions to the intensity of the conflict in an economic framework.

By use of this model, the deployment schedule for various scenarios can be presented as an input requirement for more detailed cost-effectiveness studies of alternative transportation means as represented by the linear programming Mobility and Transportation Resources Allocation Model. This will avoid the arbitrary specification of a deployment force closure schedule. Furthermore, the game theory model provides as an output the weighting of the deployment, resupply, and facilities support missions as a function of time, and thus should improve the validity of these cost-effectiveness studies.

The basic significance of this model is that it provides quantitative criteria for rapid deployment. This permits the quick assessment of hundreds of possible scenarios for the applicability of a rapid deployment policy. Other studies which have attempted to determine the value of rapid deployment have not really done so because absence of the restriction of a finite logistics pipeline constraint forces the result in favor of rapid deployment. The "mostest" in "getting there fustest with the mostest" implies no pipeline restrictions. When these restrictions exist, investment of resources in pipeline buildup rather than in rapidly deployed forces may become an optimal policy.

On the other hand, when rapid deployment is an optimal policy, the value of time is clearly given by the model. An exercise scenario was run in which the deployment for 3 days in a 30-day battle was reduced by 20 percent in order to ship facilities support cargo to build two airfields at the beginning of the campaign. In a five-division battle, this non-optimal strategy caused the Blue player to suffer 8 percent greater attrition than from an optimal strategy and to use 3.7 percent more total deployed forces for the same outcome, this despite the availability of larger pipeline capacity from the use of two additional airfields. When time is of the essence, it really is.

This model represents the first attempt to apply the powerful tools of game theory to problems of logistic pipeline allocation and rapid deployment. The point of view of the theory of games has been missed in most past airlift and sealift studies. It is proposed by this model to incorporate it and thus bring to bear a better appraisal of the influences of competitive factors in a hostile environment. Although this paper does not completely solve this game by calculating optimal trajectory state space and determining optimal strategies as a function of time and position, the incomplete solution herein presented is felt to be significant in formulating a new point of view for military logistics decisions.

REFERENCES

[1] Engel, J. H., "A Verification of Lanchester's Law," Operations Research **2,** 163–171 (May 1954).

[2] Fain, W. W., J. B. Fain, and H. W. Karr, "A Tactical Warfare Simulation Program," Naval Research Logistics Quarterly **13,** 413–436 (Dec. 1966).

[3] Isaacs, R., *Differential Games* (John Wiley, New York, 1965).

[4] Lanchester, F. W., *Aircraft in Warfare: The Dawn of the Fourth Arm.* (Constable & Co., London, England, 1916).

[5] McNamara, R. S., Statement by the Secretary of Defense, Robert S. McNamara, on the Fiscal Year 1969–1973 Defense Program and the 1969 Defense Budget, 22 January 1968.

[6] Weiss, H. K., "Combat Models and Historical Data: The U.S. Civil War," Operations Research **14,** 759–790 (Sept.–Oct. 1966).

[7] Wentling, L. G., Jr., G. R. Fitzpatrick, M. J. O'Brien, and J. C. Whiton, "A Programming Model for the Design of Strategic Deployment Systems," Research Analysis Corporation Technical Paper RAC–TP–211 (Sept. 1966).

MULTIPLE DISPATCHES IN A POISSON PROCESS

Colin E. Bell

University of California at Irvine
Irvine, California

ABSTRACT

An optimal policy is characterized for operating the following system. Customers arrive in $[O, T]$ according to a homogeneous Poisson process. Instantaneous services are provided at times O and T. Additional instantaneous services can be provided at N intermediate stopping times. These times must be chosen to minimize the total expected customer-hours in $[O, T]$ spent waiting for service.

The purpose of this paper is to extend the results of Ross [3]. Consider a queuing system which is empty at time O. Customers arrive according to a homogeneous poisson process with rate, λ. All customers present at the fixed time, T, are served immediately. T is the end of the time horizon. Additional services can be provided at stopping times $\tau_1, \tau_2, \ldots, \tau_N$ with $P(O \leqslant \tau_1 \leqslant \tau_2 \leqslant \ldots \leqslant \tau_N \leqslant T) = 1$. All customers present at these stopping times are served instantaneously. The problem is to determine a set of optimal stopping times $\{\tau_i^*, i = 1, \ldots, N\}$ which minimizes the expected number of customer-hours spent waiting for service (hereafter called expected wait). We shall refer to this problem as the N-dispatch problem with horizon T.

For $N = 1$, Ross [3] has shown that τ_1^* exists and is given by (1),

$$\tau_1^* = \min \{t | X_t \geqslant \lambda(T - t)\}, \tag{1}$$

where X_t is the number of customers present at time, t. In words, the rule is "stop (dispatch) at the first moment when the number of customers present is greater than or equal to the expected number of arrivals in the remainder of the time horizon."

Although Ross' result was obtained directly in [3], it can also be derived by applying Ross' [2] continuous analogue of Chow and Robbins' [1] monotone case. We shall use this latter approach for N greater than 1. We first introduce some convenient preliminaries.

For the N dispatch problem with horizon T, we define:

$W_N(T) \equiv$ infimum over all policies of the expected wait, $N \geqslant 0$

$W_N(T) \equiv 0$ if $T \leqslant 0, \qquad N \geqslant 0$

$W_N'^{-} \equiv$ left derivative with respect to T.

$z_N(T) \equiv$ expected total number of customers dispatched at the optimal stopping times $\tau_1^*, \ldots, \tau_N^*$

$z_0(\cdot) \equiv 0$

$X_t \equiv$ number of customer arrivals in the time interval $[0, t]$.

We shall eventually show that

$$\tau_1^* = \min \{t | X_t \geq \lambda(T-t) - z_{N-1}(T-t)\}, \tag{2}$$

where $z_{N-1}(T-t)$ in (2) is the expected number of customers dispatched at the optimal stopping times in the $N-1$ dispatch problem with horizon $T-t$.

First, we shall need

LEMMA 1: $W_N(T)$ *is convex in* T.

PROOF: By induction.

1. For $N=0$: $W_0(T) = \lambda T^2/2$, $T > 0$
$= 0$, $T \leq 0$,

clearly $W_0(T)$ is convex.

2. For N, assuming $W_{N-1}(T)$ is convex:

For any stopping time, τ, let $S(\tau)$ be the expected wait in the time interval $[0, \tau]$. Thus

$$S(\tau) = E_\tau \left[\int_0^\tau X_t \, dt \right].$$

For $0 < \epsilon < T$, consider time horizons $T-\epsilon$, T, and $T+\epsilon$. Let τ_- and τ_+ be first stopping times in the N dispatch problems with horizons $T-\epsilon$ and $T+\epsilon$, respectively. Let $\tau = \min(\tau_+, T)$ and let τ_1 be the stopping time for horizon T that is determined by flipping a fair coin at time 0 and choosing between τ_- and τ according as heads or tails come up.

We shall require two facts relating τ to τ_+:

Since $W_{N-1}(x) = 0$ for non-positive x,

$$EW_{N-1}(T-\tau) = EW_{N-1}(T-\tau_+), \tag{3}$$

and since $P(\tau_+ \geq \tau) = 1$,

$$S(\tau_+) \geq S(\tau). \tag{4}$$

Since $W_N(T)$ is the infimum over all policies of the expected wait:

$$W_N(T) \leq S(\tau_1) + EW_{N-1}(T-\tau_1) = \tfrac{1}{2}\{S(\tau) + EW_{N-1}(T-\tau)\} + \tfrac{1}{2}\{S(\tau_-) + EW_{N-1}(T-\tau_-)\}$$

Relations (3) and (4) then yield

$$W_N(T) \leq \tfrac{1}{2}\{S(\tau_+) + EW_{N-1}(T-\tau_+)\} + \tfrac{1}{2}\{S(\tau_-) + EW_{N-1}(T-\tau_-)\}, \tag{5}$$

and convexity of W_{N-1} permits the transition from (5) to

$$W_N(T) \leq \tfrac{1}{2}\{S(\tau_+) + EW_{N-1}(T+\epsilon-\tau_+)\} + \tfrac{1}{2}\{S(\tau_-) + EW_{N-1}(T-\epsilon-\tau_-)\}. \tag{6}$$

Since τ_+ and τ_- are arbitrary first stopping times for their respective horizons, it follows immediately that

$$W_N(T) \leq \tfrac{1}{2}\{W_N(T+\epsilon) + W_N(T-\epsilon)\}.$$

(Q.E.D.)

Our objective is to choose τ_1^* to minimize

$$E\left\{\int_0^\tau X_s ds + W_{N-1}(T-\tau)\right\}$$

which we can write as

$$W_{N-1}(T) + E\left\{\int_0^\tau X_s - W'^{-}_{N-1}(T-s)\,ds\right\},$$

due to the convexity of W_{N-1} (Lemma 1).

Let $B=\{(x, t)\colon x - W'^{-}_{N-1}(T-t) \geq 0\}$. It is trivial to show that B is a closed set, that is

$$(x_t, t)\epsilon B \Rightarrow P[(X_{t+\delta}, t+\delta)\epsilon B] = 1$$

for all stopping times δ with $P(O \leq \delta \leq T-t) = 1$. Ross' continuous monotone case can then be applied to establish the optimality of τ_1^* in (7):

$$\tau_1^* = \inf\{t \geq 0: \quad (x_t, t)\epsilon B\}. \tag{7}$$

To establish our desired result (2), we need only prove Lemma 2:

LEMMA 2: $W'^{-}_{N-1}(t) = \lambda t - z_{N-1}(t)$ for all $t > 0$.

PROOF:

Let $\tau', \ldots, \tau'_{N-1}$ be optimal stopping times for horizon t. For $\epsilon > 0$, let $\tau_1^\epsilon, \ldots, \tau_{N-1}^\epsilon$ be optimal stopping times for horizon $t-\epsilon$. Let $\tau_i = \min\{\tau_i', t-\epsilon\}$, $i=1, \ldots, N-1$.

1. $W_{N-1}(t)$ does not exceed the expected wait from any policy in the t horizon problem. Consider the policy determined by $\tau_1^\epsilon, \ldots, \tau_{N-1}^\epsilon$. Then

$$W_{N-1}(t) \leq W_{N-1}(t-\epsilon) + \epsilon[\lambda(t-\epsilon) - z_{N-1}(t-\epsilon)] + o(\epsilon), \tag{8}$$

where the second term on the right hand side of (8) represents the expected wait in time $(t-\epsilon, t]$ of those present at time $t-\epsilon$ and $o(\epsilon)$ represents the expected wait of new arrivals in $(t-\epsilon, t]$. Clearly (8) implies

$$W'^{-}_{N-1}(t) \leq \lambda t - z_{N-1}(t).$$

2. Consider the time horizon $t-\epsilon$. $W_{N-1}(t-\epsilon)$ does not exceed the expected wait from any policy. Let τ_i, $i=1, \ldots, N-1$ determine the policy to be considered. The expected wait associated with this policy can then be found by substracting the expected wait in $[t-\epsilon, t]$ from the expected wait in $[0, t]$.

The expected wait in $[0, t]$ under $\tau_1, \tau_2, \ldots, \tau_{N-1}$ is no greater than

$$W_{N-1}(t) + \frac{\lambda\epsilon^2}{2} = W_{N-1}(t) + o(\epsilon).$$

Dispatching at τ_i instead of τ_i' can increase the waiting time of only those customers who arrive in $(t-\epsilon, t)$. This expected increase can not exceed $\lambda\epsilon^2/2$.

The expected wait in $(t-\epsilon, t)$ under t_i is at least $\epsilon[\lambda(t-\epsilon)-z_{N-1}(t)]$ since the expected number present at time $t-\epsilon$ is at least $\lambda(t-\epsilon)-z_{N-1}(t)$. Thus the expected wait in $[o, t-\epsilon]$ is less than or equal to

$$W_{N-1}(t)+o(\epsilon)-\epsilon[\lambda(t-\epsilon)-z_{N-1}(t)].$$

But $W_{N-1}(t-\epsilon) \leq W_{N-1}(t)+o(\epsilon)-\epsilon[\lambda(t-\epsilon)-z_{N-1}(t)]$. It thus follows that

$$W'^{-}_{N-1}(t) \geq \lambda t - z_{N-1}(t). \qquad \text{Q.E.D.}$$

For $N \geq 1$, optimal stopping times can be generated recursively by (2). Assuming that optimal times are known for the $N-1$ dispatch problems with time horizons between 0 and T, the initial dispatch time τ_1^* has the following interpretation: "stop at the first moment, t, when the number of customers present is greater than or equal to the expected number present at the end of the time horizon using the optimal rules in the $N-1$ dispatch problem with horizon $T-t$."

The optimal first stopping time, τ_1^*, is difficult to compute since knowledge of $z_{N-1}(t)$ is required for all t between 0 and T. A logical avenue for future research is to investigate the behavior of suboptimal yet computationally easy rules. One such set of rules comes immediately to mind. Under the optimal rules, we would expect that the proportion of customers dispatched at the $N-1$ optional stopping times would be roughly $(N-1)/N$ and that the proportion dispatched at the fixed time, T, would be roughly $1/N$. Thus $z_{N-1}(t)$ would be close to $(N-1)\lambda t/N$ and $\lambda t - z_{N-1}(t)$ close to $\lambda t/N$. The degree of suboptimality of the rules generated recursively by

$$\tau_1 = \min\{t | X_t \geq \lambda(T-t)/N\}$$

is an open question for future research.

REFERENCES

[1] Chow, Y. S. and H. Robbins, "A Martingale Systems Theorem and Applications," *Proceedings of the Fourth Berkeley Symposim on Mathematical Statistics and Probability* (University of California Press, 1961).

[2] Ross, S. M. "Infinitessimal Look Ahead Stopping Rules," to appear in Annals of Mathematical Statistics.

[3] Ross, S. M., "Optimal Dispatching of a Posson Process," Journal of Applied Probability, December 1969, pp. 602–609.

* * *

A COMPARATIVE STUDY OF DEMAND FORECASTING TECHNIQUES FOR MILITARY HELICOPTER SPARE PARTS*

Robert E. Markland

University of Missouri
St. Louis, Missouri

ABSTRACT

This paper deals with techniques applicable to predicting spare parts demand for military helicopters. The military helicopter is a distinct weapons system, whose unique configuration may preclude the direct application of forecasting techniques which have proved successful for other weapon systems. Furthermore, although the military helicopter has become extremely important tactically in modern warfare, it has received scant attention in terms of research concerning its supply support.

Specifically, this paper summarizes research done to measure and compare the forecasting accuracy of six mathematical models, as they were applied to three prominent military helicopters. In addition, the paper describes attempts that were made to define, where possible, the conditions under which a specific forecasting technique might be applicable. In general, it is shown that the most accurate set of helicopter spare parts demand forecasts are produced by a second order polynomial exponential smoothing model. This model is observed to have most accurately described the highly volatile, and upward-trended demand time series which were the subject of the study.

INTRODUCTION

An efficient method for determining the demand for military equipment spare parts is critical to the design of an effective and economical logistics system. Demand forecasts are of importance to the logistics system manager in provisioning, budgeting, procurement, repair scheduling, determining distribution and stockage policies, and in other decision areas. Obviously, the operating cost of the logistics system and its number of supply failures is significantly affected by the accuracy with which demand is predicted. Because forecasts of demand cannot be perfect, the logistics system manager is confronted with making decisions under conditions of demand uncertainty. Any technique or procedure which reduces this uncertainty increases the probability of improvement in the aggregate logistic decision process. It is upon this phase of logistics management—the forecasting of spare parts demand—that this study has been conducted.

This particular study was concerned with the analysis, design, and testing of statistical forecasting systems applicable to spare parts demand for United States Army helicopters. The data used in the study applied to those helicopters under the control of the United States Army Aviation Systems Command (AVSCOM), St. Louis, Missouri.

Research concerning time series forecasting for military equipment spare parts has been extensive. One significant portion of this previous research has been concerned with the cost savings resulting from improved demand prediction [13, 14, 16, 34–38]. Another important research topic has

*The author would like to acknowledge the assistance of the Aviation Systems Command, St. Louis, Missouri in providing data. The computer centers of Washington University and University of Missouri, St. Louis, Missouri, provided computer running time for the study.

been that of spare parts demand prediction for fixed wing aircraft and missles [3, 6, 7, 10, 18, 32, 33]. Research concerning spare parts demand prediction for rotary winged aircraft (helicopters) has been extremely limited; primary concern has been in statistically analyzing failure data [5, 31, 39].

It should be emphasized that the unique configuration of a rotary winged aircraft may well preclude the direct application of forecasting techniques which have proved successful for other weapon systems. In essence, this question was one of the major considerations of this study. The importance of the helicopter as distinct from fixed wing aircraft in Army aviation should also be emphasized. During the period of this study, almost 80 percent of the aircraft being supported by AVSCOM were helicopters [40]. This proportion could be expected to increase as the Army continues to emphasize increased air mobility in ground combat. Finally, it may also be expected that the increased reliance on helicopters for combat operations will increase the complexity of demand prediction, because past experience has indicated that helicopters require more logistical support than do fixed wing aircraft [41].

REVIEW OF PREVIOUS DEMAND PREDICTION RESEARCH

The major emphasis of previous research concerning military spares demand prediction has been on isolating a "preferred" technique for demand prediction. The demand forecasting techniques previously explored can be summarized into eight general categories, as follows:

1. *Regression Techniques.* Regression techniques are well developed and have been widely applied to time series forecasting. Applications of such techniques to military spares demand prediction have usually involved an attempt at developing a relationship between spares demand (dependent variable) and various activity measures (independent variables), such as flying hours, calendar time, number of sorties flown, and engine operating hours [3, 11, 19].

2. *Issue Interval Techniques.* Use of an issue interval technique is based upon the assumption of a constant demand rate over time, so that the relationship between past and future demand populations is a function only of total program activity in each period. The total demand for a given part during an experience period is divided by the total activity of the aircraft during the experience period to give an average forecast rate. The forecast for a future period is then obtained by multiplying the average forecast rate by the planned activity for the future period [1–3].

3. *Service Life Techniques.* Use of service life techniques require estimates of the service life characteristics of the part, derived from historical data. Future spares demands are then assumed to be affected by both the planned future activity for the weapons system, and by the service life characteristic of the part [2, 20, 23, 24].

4. *Exponential Smoothing Techniques.* Use of exponential smoothing techniques does not require the assumption of dependence of demand for a part on the level of activity of the weapons system. Instead, forecasted future demand is based solely on the exponentially weighted previous demand data [3, 4].

5. *Bayesian Techniques.* Use of a Bayesian technique involves combining a priori information (for example, the initial estimate of demand provided by an engineer familiar with the technical characteristics of the part) and data derived from subsequent events (early demand experience), to predict future demand. As operational experience grows there is a gradual shift, by means of Bayesian revision, from the initial estimate of the demand rate to the actual demand rate being experienced [5, 25, 26].

6. *Conditional Probability Techniques.* Use of these techniques require the empirical development of formulae for predicting the probability that a demand of Y units of an item will occur in a specified future time period, given that X units have been demanded in a past period. Conditional probabilities employing a number of probability distributions have been developed [43].

7. *Distribution Fitting Techniques.* These techniques are applicable when it is possible statistically to fit past demand data to particular probability distributions. Distribution fitting techniques have often been used in situations requiring relatively inexpensive predictions [6, 12, 22, 28].

8. *Spectral Analysis Techniques.* Spectral analysis, an outgrowth of the theory of stationary time series, attempts to decompose demand into its three basic components: trend, cyclical, and random. The significance of each of these components in the demand process is then evaluated, and the power spectrum of the underlying demand generating process is identified and used to predict future demand [3, 17].

It should be noted that these eight techniques represent a broad spectrum in terms of mathematical sophistication, ranging from the naive issue interval technique to the complex spectral analysis technique. Unfortunately, very little research concerning the comparative effectiveness of alternative forecasting techniques has been undertaken. In essence, only one major study of such a comparative nature has been made, that being a study by Astrachan, Brown, and Houghten [2]. However, even this study considered only two major types of forecasting techniques, issue interval and service life. In summary, previous military spares demand prediction research has centered almost exclusively on the development of a single preferred forecasting technique.

OBJECTIVES OF THE PRESENT STUDY

The major objectives of the present study were markedly different from those of previous studies. The first objective was to appraise critically the predictive accuracy of a number of demand prediction techniques, as they were applied to a common sample of military helicopter spare parts. The second objective of the study was to determine if there were isolatable conditions, or characteristics of helicopters and/or their associated parts which caused certain forecasting techniques to predict more accurately. The third objective was to appraise critically the forecasting technique employed by AVSCOM (issue interval technique) during the time period covered by the study. The evaluations made in the study were essentially comparative rather than absolute, and were made for a weapons system which had previously received little attention.

DATA USED IN THE STUDY

During the period of this study, AVSCOM provided spare parts support for 13 helicopter systems, 3 of which were chosen for analysis based on the following factors:

1. High helicopter density, that is, a large number of helicopters deployed to field.
2. High helicopter usage, that is, those helicopters being flown a large number of hours each month.
3. Availability of data, that is, those helicopters for which accurate demand and activity data could be obtained.

The helicopters considered as a result of the three previous factors were:

1. *The Bell UH–1D "Iroquois."* The UH–1D is a single engine, single main rotor helicopter used as the standard utility-tactical helicopter of the U.S. Army [21].

2. *The Hiller OH–23D "Raven."* The OH–23D is a single engine, single main rotor helicopter used for observation, reconnaissance, training, and medical evacuation [21].

3. *The Boeing CH–47A "Chinook."* The CH–47A is a twin turbine, tandem rotor helicopter used to transport troops, cargo, and weapons within a combat area [21].

Part samples for each of these three helicopter systems were next selected, with the samples being

confined to critical item parts in the "High-Value" or "Super High-Value" categories.* At the time of this study, there were approximately 15 critical item spare parts on the OH–23D, 25 critical item spare parts on the UH–1D, and 35 critical item spare parts on the CH–47A. A simple random sampling procedure was employed in general; however, this procedure was modified in some instances so that parts common to all three helicopter systems could be included. This was done so that certain relationships between helicopter systems could be examined. The sampling procedure resulted in samples whose parts exhibited widely varying demand characteristics. For example, the samples included inexpensive parts with very high demand rates, and expensive parts with very low demand rates. The helicopter systems and the associated parts for which demand histories were collected are shown in Table 1.

The demand data for the parts shown in Table 1 were obtained from the AVSCOM Demand Data History File. Only recurring demands, which could be expected to occur routinely as a result of helicopter activity, were considered in this study. Thus, initial support requirements, mobilization requirements, and non-recurring base requirements were eliminated from consideration. The CH–47A helicopter and the UH–1D helicopters were both used in the United States and in Vietnam. Consequently, two demand time series were collected for each of the parts in the sample for these two helicopters. In addition to demand data, helicopter activity data, in terms of flying hours and number of helicopters deployed, were collected for the same time periods for which demand histories were accumulated. In general, about 60 months of demand and activity data were available.

FORECASTING TECHNIQUES TESTED

Six forecasting methods were considered in the study, they have been designated as Prediction Techniques 1 through 6. These six techniques were selected from the eight general categories of

*Critical item spare parts were defined as those parts that were essential to the operation of the helicopter. The "High-Value" or "Super-High Value" categories included parts whose yearly issues were in excess of $25,000.

TABLE 1. *Helicopter Systems and Parts Sampled*

Parts in sample	Helicopter systems		
	UH–1D	OH–23D	CH–47A[a]
Driveshaft	X	X	X
Engine	X	X	X
Fuel Control	X	X	X
Gear Box	X	X	.....
Main Rotor Blade	X	X	X (Forward)
Main Rotor Blade	.....	.....	X (After)
Main Rotor Hub	X	X	X (After)
Scissors	X	X	.....
Swashplate	X	X	X (Forward)
Swashplate	.....	.....	X (After)
Tail Rotor Blade	X	X	.....
Transmission	X	X	X (Forward)
Transmission	.....	.....	X (After)
Total Number of Parts Sampled	10	10	10

[a] The CH–47A helicopter is powered by tandem engines. Thus, a number of its parts occur in pairs, that is, a forward and an after component corresponding to the forward and after power unit.

demand prediction techniques which were reviewed in an earlier section of the paper. A categorization of the forecasting techniques used in this study is as follows:

Category	*Forecasting Technique(s)*
Issue Interval	Prediction Technique #1
Regression	Prediction Technique #2
Exponential Smoothing	Prediction Techniques #3, #4, #5, and #6

Each prediction technique employed in this study will be briefly described at this point in the paper.

Prediction Technique #1: Issue Interval*

The total number of demands during the experience period is divided by the number of flying hours for the experience period to give an average demand rate. The rate thus obtained is multiplied by the planned number of flying hours for the prediction period to obtain the predicted number of demands for the prediction period:

$$\hat{d}_{t+i} = (D/F)\hat{f}_{t+i}, \tag{1}$$

where $\hat{d}_{t+i}$ = predicted number of demands for the $t+i$ future period;

D = observed number of demands during the experience period;

F = observed flying hours during the experience period; and

$\hat{f}_{t+i}$ = predicted flying hours for the $t+i$ future period.

Prediction Technique #2: Moving Regression

Multiple regression relationships were derived for the initial experience period, with number of demands as the dependent variable. The independent variables considered were the actual flying hours over time, helicopter density over time, and various transformations of these two variables. From 18 to 36 monthly observations were employed in the regression fitting process. Both linear and curvilinear regression equations were considered.

The "best fitting" regression equations were then used to make forecasts over the prediction period.† After each forecasting cycle, the parameters of the regression equation were revised, and forecasts for the next cycle were made. The moving regression procedure thus attempted to track the demand process over time.

Prediction Technique #3: General Exponential Smoothing

A general exponential smoothing model, employing a linear trend and a multiplicative seasonal, and based on the work of Peter R. Winters was used for forecasting [42]. The forecasting equations for the model were given by

$$\hat{d}_t = A d_t / S_{t-L} + (1-A)(\hat{d}_{t-1} + T_{t-1}) \tag{2}$$

---

*The issue interval technique was the basic demand forecasting technique employed by AVSCOM during the course of this study.

†Best fitting regression equations were generally determined by maximizing the coefficient of multiple determination. However, the significance of the net regression coefficients was also tested, and non-significant independent variables were eliminated from the multiple regression equation.

$$\hat{d}_{t+i} = (\hat{d}_t + iT_t)S_{t-L+1} \qquad i = 1, 2, \ldots, L, \tag{3}$$

where

d_t = actual demand during the t^{th} period;
$\hat{d}_t$ = forecasted demand during the t^{th} period;
$\hat{d}_{t+i}$ = forecasted demand during the $t+i$ future period;
S_{t-L} = the multiplicative seasonal effect;
T_{t-L} = the current estimate of the linear trend effect;
L = the number of months in a cycle period, the periodicity of the seasonal effect; and
A = the exponential smoothing coefficient, $0 \leq A \leq 1.0$.

Initial estimates of demand rate, the seasonal effect in the demand rate, and the trend effect in the demand rate were made using the 18 to 36 months of data in the initial experience period. Forecasts were then made over the forecasting horizon, and the trend and seasonal factors were revised using the following relationships:

$$S_t = Bd_t/\hat{d}_t + (1-B)S_{t-L} \tag{4}$$

and

$$T_t = C(\hat{d}_t - \hat{d}_{t-1}) + (1-C)T_{t-1}, \tag{5}$$

where

B = The seasonal effect coefficient, $0 \leq B \leq 1.0$; and

C = The trend effect coefficient, $0 \leq C \leq 1.0$.

These forecasts were a function of the set of coefficients, A, B, and C. Determination of the set of coefficients which produced the most accurate demand forecast was accomplished by a three-dimensional grid search procedure. A grid of values, composed of all combinations of A, B, and C from 0.0 to 0.7 in increments of 0.1 was employed. The set of A, B, and C which was found to produce the forecast having the smallest forecasting error, was selected as being optimum.

Prediction Techniques #4: Single Exponential Smoothing

A simple algebraic model was used to describe the stochastic process underlying the demand time series. Forecasts for future time periods were obtained by the relationship:

$$\hat{d}_{t+i} = \hat{a}_t, \tag{6}$$

where $\hat{d}_{t+i}$ = an estimate of demand in time period $t+i$ made in time period t; and
$\hat{a}_t$ = the simple average demand during the time periods from 0 to t.

Prediction Technique #5: Double Exponential Smoothing

A first degree polynomial model was used to describe demand over time, assuming that there was

a significant linear trend to demand. Forecasts for future periods were obtained by the relationships:

$$\hat{d}_{t+i} = \hat{a}_t + \hat{b}_t i, \tag{7}$$

where $\hat{b}_t$ = the linear trend during the time period 0 to t, and $\hat{a}_t$ is the same as previously defined.

Prediction Technique #6: Triple Exponential Smoothing

A second degree polynomial model was used to describe demand over time, assuming that there was trend in demand which was changing as a quadratic function of time. Forecasts for future periods were obtained by the relationships:

$$\hat{d}_{t+i} = \hat{a}_t + \hat{b}_t i + 1/2\hat{c}_t i^2,$$

where $\hat{c}_t$ = the rate of change of trend during time period 0 to t, and $\hat{a}_t$ and $\hat{b}_t$ are the same as previously defined.

Brown and Meyer have rigorously proved that it is possible to estimate the $(n+1)$ coefficients in an n^{th} order polynomial exponential smoothing model by a linear combination of the first $(n+1)$ orders of exponential smoothing [9]. Brown has applied the results of this proof to the derivation of equations for the coefficients $\hat{a}_t$, $\hat{b}_t$, and $\hat{c}_t$ [13]. His procedure is recursive, and requires only the specification of a smoothing coefficient A, and an estimate of the demand rate, $\hat{d}_t$. The estimate of the demand rate was once again made using the 18 to 36 months of data in the initial experience period. The smoothing coefficient, A, was then defined and forecasts were made over the forecasting horizon. These forecasts were a function of the value of the smoothing coefficient, A. Determination of the value of A which produced the most accurate demand forecast was accomplished by a one-dimensional grid search procedure. A grid of values for A, composed of all values from 0.00 to 0.50 in increments of 0.05, was employed. The value of A which was found to produce the forecast having the smallest forecasting error, was selected as being optimum. This procedure was utilized for each of the latter three exponential smoothing models.

METHODOLOGY USED IN THE STUDY

The methodology used in the study encompassed three major steps. First, a period of time known as the *initial experience period* was established. The duration of the initial experience period was determined as a function of the total length of the time series being analyzed, with a longer time series resulting in a longer initial experience period. Thus, from 18 to 36 months of data were available for the initial experience period, with the variation being dependent upon the length of the time series being analyzed. During the initial experience period, demand and activity data were collected, and initial parameters for the various forecasting models were derived.

Second, a forecast *cycle period* in months was established. The length of the cycle period was determined as a function of two factors:

1. The duration of the difference between the total length of time series and the length of the experience period. This period of time, in months, was defined as the *forecast horizon*.

2. The actual replenishment lead time used at AVSCOM. During this study, the replenishment lead time was in the range of 6 to 12 months. In essence, the cycle period was selected as being either

6 or 12 months in length, with the shorter cycle period corresponding to the shorter forecast horizon, and vice versa.

Third, a forecast was made for the first cycle period using the initialization parameters established with data from the initial experience period. At the end of this cycle period, a new set of model parameters was computed based on a new experience period which included the most recent cycle period, and, in general, excluded data from the early part of the previous experience period. Then a new forecast for the next cycle period was made, and the previous parameter revision process was repeated. This procedure continued over the entire forecasting horizon.

MEASUREMENT OF FORECASTING ERROR

Upon completion of the forecasting procedure, forecasting errors were computed for the series of L month cycle period forecasts. The forecast error, $e_{t,L}$, for an L month cycle was given by:

$$e_{t,L}=\sum_{i=1}^{L}(\hat{d}_{t+i}-d_{t+i}) \qquad t=1, 2, \ldots, n-L+1 \tag{9}$$

where

$\hat{d}_{t+i}$ = forecasted demand in month $t+i$;

d_{t+i} = actual demand in month $t+i$;

L = number of months in a cycle period (6 or 12); and

n = number of months in the forecasting horizon.

Since forecasts were made in L month cycles over the forecasting horizon of n months, there were $k=n-L+1$ such summations. Each summation produced a total forecasting error for one cycle. The standard deviation, δ_L, for k such L month cycles was given by

$$\delta_L=\sqrt{\sum_{t=1}^{k}\frac{(e_{t,L})^2}{k-1}}. \tag{10}$$

Finally, the coefficient of variation, CV_L, over the forecasting horizon was given by

$$CV_L=\frac{\delta_L}{\sum_{i=1}^{L} d_{t+i}/k} \qquad t=1, 2, \ldots, n-L+1. \tag{11}$$

The coefficient of variation over the forecasting horizon was used as the primary measure of forecast accuracy.

Bias was measured by computing the relative cycle error, RE_L, given by

$$RE_L=\frac{\sum_{i=1}^{L}(\hat{d}_{t+i}-d_{t+i})}{\sum_{i=1}^{L}(d_{t+i})} \qquad t=1, 2, \ldots, n-L+1. \tag{12}$$

The measurement of relative bias over the forecasting horizon was used to compare the underprediction and overprediction resulting from the application of the various forecasting techniques. Underpredic-

tions were inherently considered to be more undesirable than overpredictions, since they prevented the helicopter from accomplishing its mission.

TEST RESULTS

The coefficient of variation and relative errors for each of the three helicopters and for each of their associated parts sample are shown in Tables A–1 through A–5 in the appendix. In order to summarize these extensive results, mean coefficients of variation and mean relative errors were computed for each helicopter and type of demand. This summary is presented in Table 2.

TABLE 2. *Summary of Test Results*

Helicopter system	Type of demand	Techniques					
		1	2	3	4	5	6
A. Mean Coefficients of Variation							
OH–23D	Non-Vietnam	0.47	0.77	0.49	0.49	0.47	0.42
CH–47A	Non-Vietnam	1.86	0.61	0.69	0.46	0.43	0.38
	Vietnam	0.39	1.14	0.44	0.38	0.34	0.30
UH–1D	Non-Vietnam	1.26	0.76	0.59	0.42	0.36	0.28
	Vietnam	0.30	0.38	0.26	0.26	0.26	0.25
B. Mean Relative Errors							
OH–23D	Non-Vietnam	−0.17	−0.64	−0.56	+0.01	+0.04	+0.05
CH–47A	Non-Vietnam	+1.57	+0.28	+0.15	−0.06	−0.05	−0.04
	Vietnam	+0.12	+0.42	+0.06	+0.17	+0.16	+0.15
UH–1D	Non-Vietnam	+1.12	+0.09	−0.02	−0.02	−0.01	+0.00
	Vietnam	−0.10	−0.22	−0.01	−0.01	−0.03	−0.02

An appraisal of the forecasting accuracy results presented in Table 2 clearly indicated the general superiority of the triple exponential smoothing technique. The lowest mean coefficient of variation was produced by the triple exponential smoothing technique for each of the five "helicopter" and "type of demand" categories. Similarly, the lowest mean relative error was produced by the triple exponential smoothing technique for two of the five "helicopter" and "type of demand" categories and was only slightly larger for the remaining three demand categories.

A comparison of the test results for a number of different parts sample classifications was next made. The following classifications were employed:

1. Helicopter mission.
2. Technical property of the parts.
3. Theater of operations.
4. Statistical property of the parts.

Mean coefficients of variation were next computed for each of the six forecasting techniques within the framework of the various parts sample classifications. A summary of the test results made in this manner is presented in Table 3. These results again strongly indicated the marked superiority of the triple exponential smoothing technique. In all categories, the smallest mean coefficient of variation was produced by triple exponential smoothing.

A final series of tests was made to compare the statistical significance of the differences between

TABLE 3. *Comparison of Test Results by Parts Sample Classifications*

(Mean Coefficients of Variation)

Classifications	Techniques					
	1	2	3	4	5	6
Helicopter Missions:						
Training	0.47	0.77	0.49	0.49	0.47	0.42
Combat—Transport	1.12	0.87	0.56	0.42	0.39	0.34
Combat—Assault	0.78	0.57	0.43	0.34	0.31	0.27
Technical Property:						
Vulnerable and/or Exposed	0.41	0.63	0.31	0.28	0.24	0.21
Power Transmission	0.63	0.65	0.39	0.34	0.31	0.29
Mandatory Replacement	0.52	0.31	0.31	0.27	0.25	0.24
Theater of Operations:						
Non-Vietnam	1.55	0.68	0.64	0.51	0.43	0.33
Vietnam	0.35	0.76	0.35	0.30	0.28	0.27
Statistical Property:[a]						
Low-Mean Dem./Month (0.00–5.00)	0.61	0.67	0.39	0.34	0.27	0.25
Medium-Mean Dem./Month (5.01–25.00)	0.63	0.73	0.36	0.33	0.30	0.29
High-Mean Dem./Month (25.01+)	0.46	0.55	0.35	0.34	0.28	0.24
Low-Coeff. of Dispersion (0.00–0.90)	0.55	0.53	0.38	0.33	0.30	0.27
High-Coeff. of Dispersion (0.91+)	0.55	0.78	0.34	0.33	0.32	0.25

[a] Class intervals for each of the statistical properties are shown in parentheses.

the mean coefficients of variation produced by the various forecasting techniques. The two sample *t*-test was employed, with the significance level for the difference between each pair of coefficients of variation being computed. The significance level was interpreted as the probability of a difference in sample means being as large as that observed, purely due to random effects. A high level of significance indicated that such a difference could be attributed to chance, while a low level of significance indicated that the observed difference was not due to chance alone. A summary of the test results made in this manner is presented in Table 4.*

The results shown in Table 4 reinforced, once again, the contention of superiority of triple exponentional smoothing. For 21 of the 25 comparisons made, there was less than a 40 percent chance that the difference in mean coefficients of variation produced by triple exponential smoothing and any of

*For sake of brevity, only comparisons involving the triple exponential smoothing model are presented here. More comprehensive results may be found in Markland [27].

TABLE 4. *Significance Levels for Differences Between Mean Coefficients of Variation*[a]

Helicopter System	Type of Demand	Techniques				
		1 vs 6	2 vs 6	3 vs 6	4 vs 6	5 vs 6
OH–23D.....	Non-Vietnam.....	0.40	0.16	0.36	0.40	0.31
CH–47A.....	Non-Vietnam.....	0.03	0.23	0.15	0.33	0.37
	Vietnam.....	0.30	0.03	0.25	0.40	0.43
UH–1D	Non-Vietnam.....	0.04	0.06	0.10	0.14	0.17
	Vietnam.....	0.36	0.22	0.95	0.95	0.95

[a] Each significance level was computed by a two sample *t*-test with 18 degrees of freedom.

the other techniques would have occurred purely due to chance. Thus, there was a very strong indication that triple exponential smoothing produced forecasts which were significantly better than those produced by any of the other forecasting techniques in a statistical sense.

DISCUSSION OF RESULTS

The objectives of this study, as stated earlier, were three-fold. The initial objective was to compare the predictive accuracy of a number of demand prediction techniques, as they were applied to a common sample of military helicopter spare parts. The findings of this study strongly indicated the superiority of the triple exponential smoothing technique among the six techniques tested.

The second objective of the study was to determine if there were isolatable conditions, or characteristics of helicopters and/or their associated parts which caused certain forecasting techniques to predict more accurately. The results of testing were largely inconclusive in this area, as the triple exponential smoothing technique produced the most accurate sets of forecasts for virtually all of the part sample classifications.

The third objective of the study was to appraise critically the forecasting procedure utilized by AVSCOM during the time period covered by the study. Fruitful results were obtained in this regard, as it was conclusively shown that the issue interval technique produced relatively poor forecasting results when compared to the other methods tested. The result had obvious implications to the agency which participated in the investigation.

In conclusion, reasons for the superiority of the triple exponential smoothing technique can be suggested. The majority of the time series considered in the investigation had volatile demand patterns, with demand over time generally subject to a substantial upward trend. In addition, month to month fluctuations in demand were often pronounced. Now, the triple exponential smoothing model was explicitly formulated to consider demand processes having a changing, and hence nonlinear, trend in demand. In addition, the updating procedure used for the parameters of the triple exponential smoothing model allowed for rapid recognition of fundamental long-run changes in demand, as well as short-run fluctuations in demand which were of a more random nature. The general exponential smoothing model and the double exponential smoothing models considered only a linear trend. As a result, their parameter updating procedures reacted more slowly to fundamental long-run changes in demand, and tended to blur short-term changes in trend. The regression equation forecasting model implicity considered the presence of a trend in demand as a part of the derivation and revision of the regression equations used to make forecasts. However, these regression equations were derived for a time period which considerably lagged the time periods for which forecasts were being made (by an average of 3 to 6 months). Similarly, in the issue interval model, the forecasting parameter, the demand rate, was derived for a time period which lagged the time periods for which forecasts were being made by an average of 3 to 6 months. The results obtained by using triple exponential smoothing suggested that some of the most important trend effects were of a short duration, that is, 3 months or less. Essentially, the regression equation model and the issue interval model ignored these short-term trend effects in their parameter updating procedures. Since short-run trend effects were generally pronounced in the time series considered in this research, very inaccurate results were obtained from these two forecasting models.

CONCLUSION

In conclusion it must be stated that the present study is in no way definitive, but is rather a consideration of the general problem of military spares demand prediction for a unique and heretofore

unexamined weapons system. For the government agency which was the subject of the research, the test results indicated that the triple exponential smoothing technique should replace the issue interval method, with a resultant increase in forecasting accuracy.

The general area of demand prediction for military equipment spares parts remains very fertile for continued study. Certainly, the important question of the cost of overestimation versus the cost of underestimation deserves increased attention. Similarly, moving regression equation forecasting, although unsuccessful in this study, offers promise if more sophisticated data for the independent variable can be derived (e.g., hours flown at high speed, combat hours, number of takeoffs and landings, etc.). Finally, the dynamic character of military equipment spare parts time series seems to suggest that an adaptive or error tracking forecasting procedure might improve forecasting accuracy.

Appendix

TABLE A–1. *Coefficients of Variation*
(OH–23D Parts Sample)

Part description	Techniques					
	1	2	3	4	5	6
Driveshaft	0.54	0.96	0.60	0.46	0.37	0.29
Engine	0.19	0.18	0.92	0.93	0.96	0.94
Fuel control	0.53	0.66	0.50	0.52	0.51	0.45
Gear box	0.34	0.53	0.35	0.37	0.29	0.23
Main rotor blade	0.35	0.39	0.19	0.20	0.27	0.29
Main rotor hub	0.75	1.25	0.74	0.76	0.68	0.52
Scissors	0.53	1.37	0.24	0.29	0.37	0.42
Swashplate	0.71	1.02	0.51	0.56	0.54	0.53
Tail rotor blade	0.42	0.91	0.43	0.32	0.29	0.18
Transmission	0.39	0.44	0.45	0.44	0.37	0.31

TABLE A–2. *Relative Errors*
(OH–23D Parts Sample)

Part description	Techniques					
	1	2	3	4	5	6
Driveshaft	−0.36	−0.88	−0.48	−0.32	−0.19	−0.11
Engine	+0.12	−0.15	+0.45	+0.21	+0.27	+0.22
Fuel control	−0.37	−0.50	−0.13	+0.40	+0.39	+0.36
Gear box	+0.18	−0.30	+0.13	−0.14	−0.09	−0.07
Main rotor blade	−0.28	−0.28	−0.02	+0.01	+0.05	+0.03
Main rotor hub	−0.21	−0.98	−0.16	+0.34	+0.28	+0.20
Scissors	−0.31	−1.27	−0.01	+0.06	+0.14	+0.18
Swashplate	−0.39	−1.22	−0.37	−0.10	−0.08	−0.09
Tail rotor blade	−0.33	−1.17	−0.03	−0.11	−0.09	−0.07
Transmission	+0.21	+0.35	+0.06	−0.29	−0.22	−0.18

TABLE A–3. *Coefficients of Variation*
(CH–47A Parts Sample)

Part description	Techniques					
	1	2	3	4	5	6
Non-Vietnam Demand Time Series						
Driveshaft	0.65	0.30	1.61	0.54	0.69	0.64
Engine	1.73	0.96	0.53	0.47	0.32	0.25
Fuel control	1.75	0.93	0.79	0.80	0.78	0.78
Main rotor bl.—aft	1.13	0.27	1.10	0.27	0.21	0.19
Main rotor bl.—fwd	0.68	0.22	0.36	0.35	0.31	0.28
Main rotor hub—aft	1.00	1.04	0.51	0.51	0.48	0.45
Swashplate—aft	5.13	0.27	0.54	0.52	0.47	0.39
Swashplate—fwd	2.89	0.25	0.77	0.47	0.39	0.26
Transmission—aft	2.45	1.33	0.48	0.46	0.42	0.37
Transmission—fwd	1.01	0.48	0.23	0.25	0.24	0.22
Vietnam Demand Time Series						
Driveshaft	0.87	1.21	0.55	0.39	0.27	0.19
Engine	0.23	0.52	0.15	0.16	0.12	0.10
Fuel control	0.22	0.75	0.94	0.73	0.62	0.41
Main rotor bl.—aft	0.34	1.07	0.15	0.20	0.32	0.38
Main rotor bl.—fwd	0.31	2.06	0.45	0.39	0.31	0.26
Main rotor hub—aft	0.39	1.26	0.40	0.40	0.39	0.36
Swashplate—aft	0.27	1.00	0.48	0.33	0.19	0.13
Swashplate—fwd	0.41	0.52	0.34	0.33	0.29	0.25
Transmission—aft	0.52	1.72	0.52	0.51	0.49	0.48
Transmission—fwd	0.34	1.31	0.41	0.40	0.40	0.39

TABLE A–4. *Relative Errors*
(CH–47A Parts Sample)

Part description	Techniques					
	1	2	3	4	5	6
Non-Vietnam Demand Time Series						
Driveshaft	+0.68	−0.27	+1.25	+0.11	+0.17	+0.13
Engine	+0.96	+0.65	−0.08	+0.12	+0.10	+0.06
Fuel control	+0.93	+0.55	−0.02	−0.20	−0.16	−0.16
Main rotor bl.—aft	+1.03	−0.23	+1.00	+0.09	+0.03	+0.00
Main rotor bl.—fwd	+0.45	+0.15	+0.06	+0.09	+0.07	+0.04
Main rotor hub—aft	+0.66	+0.52	−0.10	−0.24	−0.21	−0.17
Swashplate—aft	+4.63	−0.10	−0.24	−0.43	−0.37	−0.32
Swashplate—fwd	+3.33	−0.15	−0.60	−0.37	−0.32	−0.21
Transmission—aft	+2.15	+1.23	+0.12	+0.14	+0.11	+0.17
Transmission—fwd	+0.83	+0.42	+0.09	+0.11	+0.11	+0.10
Vietnam Demand Time Series						
Driveshaft	+0.48	+0.89	+0.46	+0.36	+0.27	+0.21
Engine	+0.02	+0.39	−0.06	+0.08	+0.05	+0.03
Fuel control	−0.18	+0.20	+0.16	+0.37	+0.31	+0.26
Main rotor bl.—aft	+0.18	+0.87	−0.01	+0.11	+0.17	−0.23
Main rotor bl.—fwd	+0.25	+1.08	+0.19	+0.19	+0.14	+0.11
Main rotor hub—aft	+0.15	+1.00	−0.17	+0.16	+0.15	+0.13
Swashplate—aft	+0.12	−0.94	+0.15	−0.12	−0.09	−0.06
Swashplate—fwd	+0.26	−0.47	−0.06	−0.04	−0.03	−0.01
Transmission—aft	−0.33	+0.82	−0.33	+0.43	+0.41	+0.40
Transmission—fwd	−0.20	+0.38	−0.20	+0.20	+0.20	+0.19

TABLE A–5. *Coefficients of Variation*
(UH–1D Parts Sample)

Part description	Techniques					
	1	2	3	4	5	6
Non-Vietnam Demand Time Series						
Driveshaft	1.52	0.44	0.25	0.25	0.24	0.23
Engine	3.77	1.09	0.77	0.51	0.32	0.15
Fuel control	0.45	0.42	1.37	0.38	0.37	0.35
Gear box	0.28	0.55	0.36	0.34	0.27	0.20
Main rotor blade	1.48	0.68	0.78	0.66	0.61	0.49
Main rotor hub	1.04	1.08	0.46	0.45	0.43	0.41
Scissors	0.76	1.04	0.65	0.51	0.43	0.26
Swashplate	1.04	0.59	0.56	0.42	0.37	0.31
Tail rotor blade	1.85	1.15	0.35	0.32	0.29	0.27
Transmission	0.36	0.57	0.37	0.31	0.27	0.12
Vietnam Demand Time Series						
Driveshaft	0.53	0.62	0.26	0.28	0.31	0.36
Engine	0.19	0.23	0.17	0.19	0.23	0.24
Fuel control	0.17	0.40	0.28	0.29	0.31	0.30
Gear box	0.27	0.24	0.18	0.16	0.17	0.17
Main rotor blade	0.44	0.51	0.19	0.16	0.15	0.13
Main rotor hub	0.41	0.35	0.35	0.31	0.24	0.17
Scissors	0.36	0.59	0.35	0.32	0.29	0.26
Swashplate	0.24	0.33	0.34	0.38	0.39	0.39
Tail rotor blade	0.13	0.26	0.29	0.30	0.31	0.33
Transmission	0.30	0.26	0.20	0.19	0.17	0.15

TABLE A–6. *Relative Errors*
(UH–1D Parts Sample)

Part description	Techniques					
	1	2	3	4	5	6
Non-Vietnam Demand Time Series						
Driveshaft	+1.36	−0.35	−0.04	−0.03	−0.02	−0.02
Engine	−3.62	+0.80	−0.08	−0.07	−0.06	−0.04
Fuel control	+0.31	+0.35	+0.80	−0.26	−0.24	−0.23
Gear box	+0.17	−0.52	−0.16	−0.12	−0.06	−0.02
Main rotor blade	+1.26	+0.26	+0.37	+0.14	+0.12	+0.06
Main rotor hub	+0.63	−0.93	−0.28	+0.20	+0.18	+0.16
Scissors	+0.63	+0.82	−0.33	−0.14	−0.09	−0.01
Swashplate	+0.66	+0.06	−0.30	+0.12	+0.10	+0.07
Tail rotor blade	+1.74	+0.81	+0.02	−0.06	−0.05	−0.03
Transmission	+0.77	−0.40	−0.15	+0.07	+0.04	+0.02
Vietnam Demand Time Series						
Driveshaft	+0.47	−0.37	+0.04	−0.06	−0.09	−0.10
Engine	−0.07	−0.18	−0.05	+0.01	+0.02	+0.00
Fuel control	+0.15	−0.36	+0.05	−0.04	−0.06	−0.05
Gear box	−0.25	−0.20	−0.06	−0.03	−0.05	−0.02
Main rotor blade	−0.40	−0.11	−0.10	+0.06	+0.06	+0.05
Main rotor hub	−0.36	−0.17	−0.01	+0.14	+0.09	+0.07
Scissors	−0.20	−0.46	+0.09	+0.03	+0.02	+0.01
Swashplate	−0.09	−0.10	−0.02	−0.12	−0.16	−0.15
Tail rotor blade	−0.04	−0.13	+0.02	−0.14	−0.17	−0.18
Transmission	−0.25	−0.16	−0.02	±0.10	+0.09	+0.07

REFERENCES

[1] Alt, Franz L., "Safety Levels in Military Inventory Management," Operations Research *10*, 786–95 (Nov.–Dec. 1962).

[2] Astrachan, Max, Bernice B. Brown, and James W. Houghten, "A Comparative Study of Prediction Techniques," Santa Monica, California: The Rand Corporation, Research Memorandum RM–2811–PR (Dec. 1961).

[3] Astrachan, Max and Albert S. Cahn, eds., "Proceedings of Rand's Demand Prediction Conference—January 25–26, 1962," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3358–PR (Jan. 1963).

[4] Astrachan, Max and Craig C. Sherbrooke, "An Empirical Test of Exponential Smoothing," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3938–PR (Mar. 1964).

[5] Bott, L. R., "Statistical Method of Predicting Fatigue Life of Helicopter Dynamic Components," American Helicopter Society Journal *9*, 30–45 (Oct. 1964).

[6] Brown, Bernice B., "Characteristics of Demand For Aircraft Spare Parts," Santa Monica, California: The Rand Corporation, Research Memorandum R–292 (July 1956).

[7] Brown, Bernice B. and Murray A. Geisler, "Analysis of the Demand Patterns for B–47 Airframe Parts at Air Base Level," Santa Monica, California: The Rand Corporation, Research Memorandum RM–1297 (July 1954).

[8] Brown, Robert G., Smoothing, Forecasting, and Prediction of Discrete Time Series (Prentice-Hall, Inc., Englewood Cliffs, N.J., 1963).

[9] Brown, Robert G. and Richard F. Meyer, "The Fundamental Theorem of Exponential Smoothing," Operations Research *9*, 673–687 (Sept.–Oct. 1965).

[10] Campbell, Harrison S., "Procurement and Management of Spares," Santa Monica, California: The Rand Corporation, Research Memorandum P–3404 (July 1966).

[11] Campbell, Harrison S., "The Relationship of Resource Demands to Airbase Operations," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3428–PR (Jan. 1963).

[12] Davis, D. J., "An Analysis of Some Failure Data," J. Am. Statist. Assoc. *47*, 113–50 (June 1952).

[13] Denicoff, Marvin, Joseph P. Fennell, and Henry Solomon, "Summary of a Method For Determining the Military Worth of Spare Parts," Naval Research Logistics Quarterly *7*, 221–34 (Sept. 1960).

[14] Feeney, George J. and Craig C. Sherbrooke, "Systems Analysis and Supply Management," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3644–PR (Apr. 1964).

[15] Feeney, George J. and Craig C. Sherbrooke, "An Objective Bayes Approach For Inventory Decisions," Santa Monica, California: The Rand Corporation, Research Memorandum RM–4362–PR (Mar. 1965).

[16] Ferguson, Allen R. and Lawrence Fisher, "Stockage Policy For Medium and Low-Cost Parts," Santa Monica, California: The Rand Corporation, Research Memorandum RM–1962 (Apr. 1958).

[17] Fishmann, George S., "A Statistical Description of Demand Patterns Evolving Through Time," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3704–PR (Aug. 1963).

[18] Giffler, Bernard, "Mathematical Solution of Parts Requirement Problems," Management Science *11*, 847–66 (July 1965).

[19] Goldman, Thomas A., "Relationship Between Program Elements and System Demand For Airframe Spare Parts," Santa Monica, California: The Rand Corporation, Research Memorandum RM–1858 (Jan. 1957).

[20] Goldman, Thomas A., "A Priori Demand Prediction—A Case Study of B–52 Airframe Parts," Santa Monica, California: The Rand Corporation, Research Memorandum RM–2088 (Jan. 1958).

[21] "Helicopter Data–OH–23D, CH–47A, UH–1D," Army Aviation Magazine *12*, 68, 77, and 80 (Aug. 16, 1967).

[22] Hines, William W. and Joseph H. Moder, "Procedure For Forecasting Aircraft Engine Maintenance Removal Quantities," J. Industrial Eng. *11*, 471–476 (Nov.–Dec. 1960).

[23] Hunter, Lawrence and Frank A. Proschan, "Replacement When Constant Failure Rate Precedes Wearout," Naval Research Logistics Quarterly *8*, 127–136 (June 1961).

[24] Kamins, Milton, "Determining Checkout Intervals For Systems Subject to Random Failures," Santa Monica, California: The Rand Corporation, Research Memorandum RM–2578 (June 1960).

[25] McGlothlin, William H., "Development of Bayesian Parameters For Spare-Parts Demand Prediction," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3699–PR (July 1963).

[26] McGlothlin, William H. and Robert Radner, "The Use of Bayesian Techniques For Predicting Spare-Parts Demand," Santa Monica, California: The Rand Corporation, Research Memorandum RM–2536 (Mar. 1960).

[27] Markland, Robert E., "An Analysis of Alternative Quantitative Techniques Used to Predict Demand For Military Helicopter Spare Parts" (unpublished D.B.A. dissertation, Washington University, 1969).

[28] Mickey, M. R., "A Method For Determining Supply Quantity For The Poisson Distribution of Demand," Santa Monica, California: The Rand Corporation, Research Memorandum P–1376 (May 1958).

[29] Morrison, Donald F. and H. A. David, "The Life Distribution of a System With Spare Components," Annals of Mathematical Statistics *31*, 1084–1094 (Dec. 1960).

[30] Muth, John F., "Optimal Properties of Exponentially Weighted Forecasts," J. Am. Statist. Assoc. *55*, 299–306 (June 1960).

[31] Nash, C. D., "Reliability Analysis of Fatigue Failure," Am. Helicopter Soc. J. *9*, 46–51 (Oct. 1964).

[32] Okun, Bernard, "Design, Test, and Evaluation of Experimental Flyaway Kit," Naval Research Logistics Quarterly *7*, *109–136* (June 1960).

[33] Perowne, L., "Operational Research and the Supply of Spare Parts," Metra *2*, *161–169* (June 1963).

[34] Petersen, James W., "Savings From Procurement Deferral With Interim Contractor Support: The Case of High Value Airframe Parts," Santa Monica, California: The Rand Corporation, Research Memorandum RM–2085 (Jan. 1958).

[35] Petersen, James W. and H. W. Nelson, "Integrated Material Management," Santa Monica, California: The Rand Corporation, Research Memorandum RM–2870–PR (Oct. 1961).

[36] Petersen, James W., H. W. Nelson, and R. M. Paulson, "The Costs and Benefits of Responsive Support Operations," Santa Monica, California: The Rand Corporation, Research Memorandum RM–3871–PR (Oct. 1962).

[37] Proschan, F., "Optimal System Supply," Naval Research Logistics Quarterly *7*, 609–646 (Sept. 1961).

[38] Rauner, R. M., "Laboratory Evaluation of Supply and Procurement Policies," Santa Monica, California: The Rand Corporation, Research Memorandum R–323 (July 1958).

[39] Schauble, J. J. and P. F. Maloney, "Approach to Helicopter Structural Reliability and Fatigue Life," Am. Helicopter Soc. J. *9*, 13–19 (Oct. 1964).

[40] *Statistical Reference Book*, U.S. Army Aviation Material Command, St. Louis, Missouri (Printed Monthly Report).

[41] *Supporting Army Aircraft Throughout the World*, U.S. Army Aviation Material Command, St. Louis, Missouri (Aug. 1965) (Printed Booklet).

[42] Winters, Peter R., "Forecasting Sales by Exponentially Weighted Moving Averages," Management Science *6*, 324–342 (June 1960).

[43] Youngs, J. W. T., Murray A. Geisler, and Bernice B. Brown, "The Prediction of Demand For Aircraft Spare Parts Using The Method of Conditional Probabilities," Santa Monica, California: The Rand Corporation, Research Memorandum RM–1413 (Jan. 1955).

* * *

STOCHASTIC DUELS WITH NONREPAIRABLE WEAPONS

N. Bhashyam

Defence Science Laboratory
Delhi 6

ABSTRACT

This paper examines the effect of limitation, regarding weapons that are likely to fail during the period of deployment, on the final outcome in a stochastic duel model. Inter-firing times as well as inter-failure times have been assumed to be exponentially distributed.

INTRODUCTION

The theory of Stochastic Duels was developed by Williams and Ancker (1963) to study combat problems as a two person duel, incorporating into the analysis microscopic details of a combat—such as, single shot kill probability, time between rounds fired, surprise, cover, concealment, etc. The details of the work done by various analysts, is given in Ancker [3].

Ancker [1] introduced the idea of ammunition limitation and studied its effects on the win probabilities. A duel model incorporating the idea of time limitation was investigated by Ancker [2].

In addition to time and ammunition limitations, there is another type of limitation which arises from the nature of weapons that duelists deploy. Modern weapons are sophisticated and complex in nature. They perform their job with utmost excellence and efficiency, but at the same time are susceptible to failure while in commission. The failed weapon has to be replaced immediately, so that the duelist can actively continue in the duel. Because the duelist may carry only a finite number of weapons with him to the duel arena, in the event of all his weapons failing before he is able to kill his opponent, he practically becomes a sitting duck. This poses a problem of providing some guideline for the duelist as to the number of weapons he should carry so as to ensure as high a probability of win as possible. The model may also approximate a situation where a large quantity of weapons can not be carried to the duel arena due to operational reasons and at the same time weapon resupply during the duel is not possible.

This paper attempts to study a duel model wherein the duelists are handicapped by limited ammunition supplies together with a limited number of weapons.

DESCRIPTION OF THE MODEL

Two duelists, A and B, are engaged in a duel. They fire at each other with the aim of securing a kill. The duelist who kills his opponent wins the duel. We assume that duelist A enters the duel with a limited stock of weapons, R, together with limited ammunition supply, N. Duelist A may deploy only one of his weapons at a time. The inter-firing times for duelist A are distributed according to exponential distribution with parameter λ. The weapons deployed by A are likely to fail while in use, and the inter-failure times are assumed to exponentially distributed with parameter μ. Weapons possessed by A, but not in use have zero failure rates. The single shot kill probability for each round is constant and is equal to p. Let q stand for the probability of miss so that $p+q=1$.

Similar assumptions hold for duelist B, whose parameters are represented by placing a star (*) as superscript.

FORMULATION AND SOLUTION

Let us define the following discrete random variables:

$N(t)$: Number of nonlethal rounds fired by A up to time "t."

$$N \geq N(t) \geq 0.$$

$\alpha(t)$: Number of weapons of A that have failed up to time "t."

$$0 \leq \alpha(t) \leq R.$$

$\theta(t)$: Number of rounds fired by A to kill B by time "t."

$$1 \leq \theta(t) \leq N.$$

Using these, we define the following state probabilities:

$$P_n^r(t) = Pr[N(t)=n,\ \alpha(t)=r \mid N(0)=\alpha(0)=0]$$
$$Q^r(t) = P_r[N(t)=N,\ \alpha(t)=r \mid N(0)=\alpha(0)=0]$$
$$Q_n(t) = P_r[N(t)=n,\ \alpha(t)=R \mid N(0)=\alpha(0)=0]$$
$$A_n^r(t) = P_r[\theta(t)=n,\ \alpha(t)=r \mid N(0)=\alpha(0)=0].$$

Obviously,

$P_n^r(t)$ = Prob (B is alive at time "t," A having consumed n rounds and r weapons up to time "t")
$Q^r(t) = P_N^r(t)$
$Q_n(t) = P_n^R(t)$
and

$A_n^r(t)$ = Prob (B is dead by time "t," A having consumed n rounds and r weapons to achieve the kill).

By continuity arguments we obtain the following set of equations:

(1) $$\left(\frac{d}{dt}+\lambda+\mu\right)P_n^r(t) = \lambda q P_{n-1}^r(t) + \mu P_n^{r-1}(t), \qquad \begin{matrix} 1 \leq r \leq R-1 \\ 1 \leq n \leq N-1 \end{matrix}$$

(2) $$\left(\frac{d}{dt}+\lambda+\mu\right)P_n^0(t) = \lambda q P_{n-1}^0(t), \qquad 1 \leq n \leq N-1$$

(3) $$\left(\frac{d}{dt}+\lambda+\mu\right)P_0^r(t) = \mu P_0^{r-1}(t), \qquad 1 \leq r \leq R-1$$

(4) $$\left(\frac{d}{dt}+\lambda+\mu\right)P_0^0(t) = 0$$

(5) $$k_n^r(t) = \frac{d}{dt}A_n^r(t) = \lambda p P_{n-1}^r(t), \qquad \begin{matrix} 0 \leq r \leq R-1 \\ 1 \leq n \leq N, \end{matrix}$$

$$E^r(t)=\frac{d}{dt}Q^r(t)=\lambda q P^r_{N-1}(t), \qquad 0\leq r\leq R-1, \tag{6}$$

$$F_n(t)=\frac{d}{dt}Q_n(t)=\mu P^{R-1}_n(t), \qquad 0\leq n\leq N-1. \tag{7}$$

Initially,

$$P^0_0(0)=1. \tag{8}$$

Taking Laplace transform and denoting the Laplace transform of the function by placing a bar as superscript, i.e., $\bar{F}(s)=\int_0^\infty \exp(-st)F(t)dt$, $Re\, s\geq 0$, Eqs. (1) to (8) become:

$$S\bar{P}^r_n(s)=\lambda q\bar{P}^r_{n-1}(s)+\mu\bar{P}^{r-1}_n(s). \tag{9}$$

$$S\bar{P}^0_n(s)=\lambda q\bar{P}^0_{n-1}(s); \tag{10}$$

$$S\bar{P}^r_0(s)=\mu\bar{P}^{r-1}_0(s); \tag{11}$$

$$\bar{K}^r_n(s)=\lambda p\bar{P}^r_{n-1}(s); \tag{12}$$

$$\bar{E}^r(s)=\lambda q\bar{P}^r_{N-1}(s); \text{ and} \tag{13}$$

$$\bar{F}_n(s)=\mu\bar{P}^{R-1}_n(s); \tag{14}$$

where

$$S=s+\lambda+\mu.$$

Equations (9) to (11) give

$$\bar{P}^R_n(s)=\binom{r+n}{n}\frac{(\lambda q)^n\mu^r}{S^{n+r+1}}. \tag{15}$$

If $\bar{\hat{H}}(s)$ be the Laplace transform of $\hat{H}(t)$, the probability for the time taken by A to kill B, then

$$\bar{\hat{H}}(s)=\sum_{r=0}^{R-1}\sum_{n=1}^{N}\bar{K}^r_n(s).$$

From (12) to (15) we get

$$\begin{aligned}\bar{\hat{H}}(s)&=\lambda p\sum_{r=0}^{R-1}\sum_{n=1}^{N}\bar{P}^r_{n-1}(s)\\&=\lambda p\sum_{n=1}^{N}\sum_{r=0}^{R-1}\binom{n+r-1}{r}\frac{(\lambda q)^{n-1}\mu^r}{s^{n+r}}\\&=\frac{\lambda p}{s}\sum_{n=1}^{N}\left(\frac{\lambda q}{s}\right)^{n-1}\sum_{r=0}^{R-1}\binom{n+r-1}{r}\left(\frac{\mu}{s}\right)^r\\&=\frac{\lambda p}{s}\sum_{n=1}^{N}\left(\frac{\lambda q}{s}\right)^{n-1}\left(1-\frac{\mu}{s}\right)^{-u}[1-I^{(R,n)}_{\mu/s}].\end{aligned}$$

Using the identity

$$\sum_{j=0}^{p}\binom{q+j-1}{j}x^j=(1-x)^{-q}[1-I_x(p+1,q)],$$

where $I_x(p, q)$ is the well tabled (Pearson [4]) Incomplete Beta Function Ratio defined as

$$I_x(p, q)=\frac{B_x(p, q)}{B(p, q)}=\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)}\int_0^x y^{p-1}(1-y)^{q-1}dy,$$

where p and q are integers.

Now

$$\bar{\hat{H}}(\mathscr{s})=\frac{\lambda p}{\mathscr{s}+\lambda}\sum_{n=1}^{N}\left(\frac{\lambda q}{\mathscr{s}+\lambda}\right)^{n-1}-\frac{\lambda p}{\lambda+\mathscr{s}}\sum_{n=1}^{N}\left(\frac{\lambda q}{\lambda+\mathscr{s}}\right)^{n-1}I_{\mu/S}^{(R,n)}$$

$$=\frac{\lambda p}{\lambda p+\mathscr{s}}\left[1-\left(\frac{\lambda q}{\lambda+\mathscr{s}}\right)^{N}\right]-\frac{p}{q}\sum_{n=1}^{N}\left(\frac{\lambda q}{\lambda+\mathscr{s}}\right)^{n}I_{\mu/S}^{(R,n)}.$$

On using the identity

$$\sum_{n=1}^{N}a^nI_x(M, n)=\frac{a}{1-a}\left[-a^NI_x(M, N)+\left(\frac{x}{1-a(1-x)}\right)^{M}I_{1-a(1-x)}(M, N)\right],$$

we get

$$\bar{\hat{H}}=\frac{\lambda p}{\lambda p+\mathscr{s}}\left[1-\left(\frac{\lambda q}{\lambda+\mathscr{s}}\right)^{N}+\left(\frac{\lambda q}{\lambda+\mathscr{s}}\right)^{N}I_{\mu/S}(R, N)-\left(\frac{\mu}{\mathscr{s}+\mu+\lambda p}\right)^{R}I_{\frac{\mathscr{s}+\mu+\lambda p}{\mathscr{s}+\lambda+\mu}}(R, N)\right] \tag{16}$$

$$=\frac{\lambda p}{\lambda p+\mathscr{s}}\left[1-\left(\frac{\lambda q}{\lambda+\mathscr{s}}\right)^{N}I_{\frac{\mathscr{s}+\lambda}{S}}(N, R)-\left(\frac{\mu}{\mathscr{s}+\mu+\lambda p}\right)^{R}I_{\frac{\mathscr{s}+\mu+\lambda p}{\mathscr{s}+\mu+\lambda}}(R, N)\right]$$

since

$$I_x(p, q)+I_{1-x}(q, p)=1.$$

Let $\bar{\phi}(\mathscr{s})$ be the Laplace transform of $\phi(t)$, the probability for the time when duelist A is not in a position to fire at his opponent either due to the failure of all his R weapons or due to exhausting the ammunition stock without success. Then

$$\bar{\phi}(\mathscr{s})=\sum_{r=0}^{R-1}\bar{E}^r(\mathscr{s})+\sum_{n=0}^{N-1}\bar{F}_n(\mathscr{s}).$$

Using (13)–(15), we get

$$\bar{\phi}(\mathscr{s})=\lambda q\sum_{r-0}^{R-1}\binom{N+r-1}{r}\frac{(\lambda q)^{N-1}\mu^r}{S^{N+r}}+\mu\sum_{n=0}^{N-1}\left\{\binom{R+n-1}{n}\frac{(\lambda q)^n\mu^{R-1}}{S^{n-R}}\right\}.$$

Proceeding as before

$$\bar{\phi}(\mathscr{s})=\left(\frac{\lambda q}{\mathscr{s}+\lambda}\right)^{N}[1-I_{\mu/S}(R, N)]+\left(\frac{\mu}{\mathscr{s}+\lambda p+\mu}\right)^{R}[1-I_{\lambda q/S}(N, R)] \tag{17}$$

$$=\left(\frac{\lambda q}{\mathscr{s}+\lambda}\right)^{N}I_{\frac{\mathscr{s}+\lambda}{S}}(N, R)+\left(\frac{\mu}{\mathscr{s}+\lambda p+\mu}\right)^{R}I_{\frac{\mathscr{s}+\lambda p+\mu}{S}}(R, N).$$

Following on similar lines, we obtain $\bar{\hat{G}}(\mathscr{s})$ Laplace transform of $\hat{G}(t)$, the probability for the time taken by B to kill A, and $\bar{\psi}(\mathscr{s})$ the Laplace transform of $\psi(t)$, the probability for the time by which duelist B ceases to be an active participant in the duel as:

$$\bar{\hat{G}}(\mathscr{s})=\frac{\lambda^*p^*}{\lambda^*p^*+\mathscr{s}}\left[1-\left(\frac{\lambda^*q^*}{\lambda^*+\mathscr{s}}\right)^{N^*}I_{\frac{\lambda^*+\mathscr{s}}{S^*}}(N^*, R^*)-\left(\frac{\mu^*}{\mathscr{s}+\mu^*+\lambda^*p^*}\right)^{R^*}\cdot I_{\frac{\mathscr{s}+\mu^*+\lambda^*p^*}{S^*}}(R^*, N^*)\right] \tag{18}$$

and

$$\bar{\psi}(s)=\left(\frac{\lambda^* q^*}{\lambda^*+s}\right)^{N^*} I_{\frac{\lambda^*+s}{S^*}}(N^*, R^*)+\left(\frac{\mu^*}{s+\mu^*+\lambda^* p^*}\right)^{R^*} I_{\frac{s+\mu^*+\lambda^* p^*}{S^*}}(R^*, N^*), \tag{19}$$

where

$$S^*=s+\lambda^*+\mu^*.$$

At this stage we assume that when the duelists are forced to stop firing, either due to lack of operable weapons or due to nonavailability of ammunition, the time taken to kill the opponent is infinite. The probability of which is $\delta(t-\infty)\bar{\phi}(0)$ for A and $\delta(t-\infty)\bar{\psi}(0)$ for B.

If $\bar{H}(s)$ be the Laplace transform of $H(t)$, the probability density for the time taken by A to kill B, then

$$\bar{H}(s)=\bar{\bar{H}}(s)+\bar{\phi}(0).$$

Similarly, if $\bar{G}(s)$ has a corresponding definition for B, then

$$\bar{G}(s)=\bar{\bar{G}}(s)+\bar{\psi}(0). \tag{21}$$

EVALUATION OF WIN PROBABILITIES

Duelist B loses if he is killed by A either before he runs out of resources or afterwards.

If $P(A)$ is the probability that A wins the duel, then

$$P(A)=\int_{t=0}^{\infty}\hat{H}(t)\int_{\tau=t}^{\infty}\hat{G}(\tau)\,d\tau dt+\bar{\psi}(0)[1-\bar{\phi}(0)]. \tag{22}$$

We know

$$\hat{H}(t)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\bar{H}}(s)\exp(st)\,ds, \tag{23}$$

where the path of integration is parallel to the imaginary axis, and with c being chosen so that the poles of $\bar{\bar{H}}(s)$ lie to the left of the line of integration and $\hat{H}(s)$ is analytic to the right of it.

From (22) and (23) we have

$$\begin{aligned}P(A)&=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\bar{H}}(s)\left[\int_{t=0}^{\infty}\exp(st)\int_{\tau=t}^{\infty}\hat{G}(\tau)\,d\tau dt\right]ds+\bar{\psi}(0)[1-\bar{\phi}(0)]\\&=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\bar{H}}(s)\bar{\hat{G}}(-s)\frac{ds}{s}-\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\bar{H}}(s)\frac{ds}{s}+\bar{\psi}(0)[1-\bar{\phi}(0)],\end{aligned} \tag{24}$$

where the last line is obtained through an integration by parts and taking

$$\lim_{t\to\infty}e^{st}\int_{\tau=t}^{\infty}\hat{G}(\tau)\,d\tau=0.$$

To evaluate the integral in (24) we choose a semi-circular contour wholly lying on the right of the imaginary axis in the complex plane as shown in Figure 1. The line is such that it separates the poles of

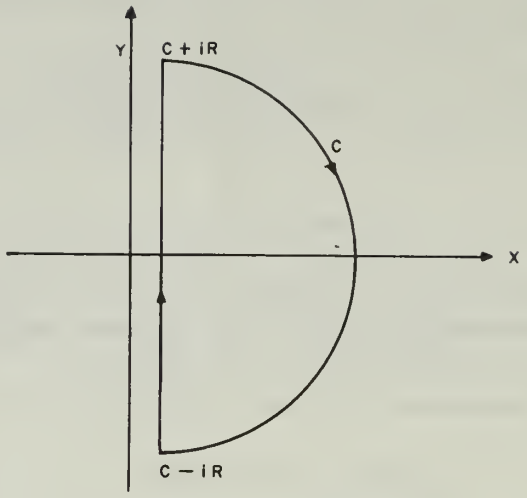


FIGURE 1.

$\bar{\hat{H}}(s)$ from those of $\bar{G}(-s)$. The poles of the intrgrand lying inside the chosen contour are those belonging to $\bar{\hat{G}}(-s)$. Hence the second integral in (24) vanishes as $\frac{1}{s}\bar{\hat{H}}(s)$ is analytic everywhere to the right of the line $c-iR$ to $c+iR$. Thus

$$P(A)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\hat{H}}(s)\bar{\hat{G}}(-s)\frac{ds}{s}+\bar{\psi}(0)[1-\bar{\phi}(0)]. \tag{25}$$

It may be remarked that as $\hat{H}(t)$ and $\hat{G}(t)$ are probability functions, therefore the integral of $\frac{1}{s}\bar{\hat{H}}(s)$ and $\frac{1}{s}\bar{\hat{H}}(s)\bar{\hat{G}}(-s)$ on C, the semi-circle (Figure 1) tend to zero as $R\to\infty$. Thus

$$P(A)=-\sum_i R_i, \tag{26}$$

where R_i is the residue at the ith pole of the integrand in (25) and summation is over all the poles lying inside the contour.

Similarly $P(B)$, the probability that B wins the duel by killing A is given by

$$P(B)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\hat{G}}(s)\bar{\hat{H}}(-s)\frac{ds}{s}+\bar{\phi}(0)[1-\bar{\psi}(0)] \tag{27}$$

$$=-\sum_j R_j^*, \tag{28}$$

where R_j^* is the residue at the jth pole of the integrand and summation is over all the poles lying in the contour in Figure 1.

$P(D)$, the probability that the duel ends in a draw is given by

$$P(D)=\bar{\psi}(0)\bar{\phi}(0). \tag{29}$$

DUEL WITH UNLIMITED AMMUNITION AND RANDOM NUMBER OF WEAPONS

We will now consider a duel situation in which both contestants have an unlimited supply of ammunition, but posess only a finite number of weapons. Let us consider that the exact number of weapons available with the duelists is not given, but the probability distribution of the same is specified. Let

$$P_r(R=m)=\alpha_m \tag{30}$$

and

$$P_r(R^*=k)=\beta_k, \tag{31}$$

such that

$$\sum_{m=0}^{\infty}\alpha_m=\sum_{k=0}^{\infty}\beta_k=1.$$

Then,

$$\bar{\bar{H}}(s)=\frac{\lambda p}{\lambda p+s}\left[1-\sum_{m=0}^{\infty}\alpha_m\left(\frac{\mu}{\mu+\lambda p+s}\right)^m\right], \tag{32}$$

$$\bar{\phi}(s)=\sum_{m=0}^{\infty}\alpha_m\left(\frac{\mu}{\mu+\lambda p+s}\right)^m, \tag{33}$$

$$\bar{\bar{G}}(s)=\frac{\lambda^* p^*}{\lambda^* p^*+s}\left[1-\sum_{k=0}^{\infty}\beta_k\left(\frac{\mu^*}{s+\mu^*+\lambda^* p^*}\right)^k\right], \tag{34}$$

and

$$\bar{\psi}(s)=\sum_{k=0}^{\infty}\beta_k\left(\frac{\mu^*}{s+\mu^*+\lambda^* p^*}\right)^k. \tag{35}$$

$P(A)$, $P(B)$, and $P(D)$, the probabilities that the duel ends in a win for A, a win for B, or a draw, are given by

$$P(A)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\bar{H}}(s)\bar{\bar{G}}(-s)\frac{ds}{s}+\sum_{k=0}^{\infty}\beta_k\left(\frac{\mu^*}{s+\mu^*+\lambda^* p^*}\right)^k\cdot\left[1-\sum_{m=0}^{\infty}\alpha_m\left(\frac{\mu}{s+\mu+\lambda p}\right)^m\right], \tag{36}$$

$$P(B)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty}\bar{\bar{G}}(s)\bar{\bar{H}}(-s)\frac{ds}{s}+\sum_{m=0}^{\infty}\alpha_m\left(\frac{\mu}{s+\mu+\lambda p}\right)^m\cdot\left[1-\sum_{k=0}^{\infty}\beta_k\left(\frac{\mu^*}{s+\mu^*+\lambda^* p^*}\right)^k\right], \tag{37}$$

and

$$P(D)=\sum_{m=0}^{\infty}\alpha_m\left(\frac{\mu}{\mu+\lambda p}\right)^m\cdot\sum_{k=0}^{\infty}\beta_k\left(\frac{\mu^*}{\lambda^* p^*+\mu^*}\right)^K. \tag{38}$$

PARTICULAR CASES

CASE 1: A has unlimited ammunition with limited weapons; B has unlimited ammunition and a failure-free weapon.

The case when A has unlimited ammunition, but limited weapons is obtained from (16) and (17) by letting $N\to\infty$ or from (32) and (33) by taking $\alpha_m=0$ for $m\neq R$ and $\alpha_m=1$ for $m=R$.

$$\bar{\tilde{H}}(\mathcal{s})=\frac{\lambda p}{\lambda p+\mathcal{s}}\left[1-\left(\frac{\mu}{\mathcal{s}+\mu+\lambda p}\right)^R\right] \tag{39}$$

and

$$\bar{\phi}(0)=\left(\frac{\mu}{\mu+\lambda p}\right)^R. \tag{40}$$

B has failure-free weapons, therefore

$$\bar{G}(\mathcal{s})=\bar{\tilde{G}}(\mathcal{s})=\frac{\lambda^* p^*}{\lambda^* p^*+\mathcal{s}} \tag{41}$$

and

$$\bar{\psi}(0)=0, \tag{42}$$

obtained from (18) and (19) by setting $\mu^*=0$ and $N^*\to\infty$ or from (34) and (35) by taking $\beta_k=0$ for $K\neq 1$, $\beta_k=1$ for $K=1$ and $\mu^*=0$. Therefore, $P(A)$ is given by

$$P(A)=\frac{\lambda p\lambda^* p^*}{2\pi i}\int_{c-i\infty}^{c+i\infty}\left[1-\left(\frac{\mu}{\mathcal{s}+\mu+\lambda p}\right)^R\right]\frac{d\mathcal{s}}{\mathcal{s}(\mathcal{s}+\lambda p)(\lambda^* p^*-\mathcal{s})}.$$

Integrating around a semi-circular contour as in Figure 1, we find the integrand has a simple pole at $\mathcal{s}=\lambda^* p^*$. Hence

$$P(A)=\frac{\lambda p}{\lambda p+\lambda^* p^*}\left[1-\left(\frac{\mu}{\mu+\lambda p+\lambda^* p^*}\right)^R\right]. \tag{43}$$

Similarly,

$$P(B)=\frac{\lambda^* p^*}{\lambda p+\lambda^* p^*}\left[1+\frac{\lambda p}{\lambda^* p^*}\left(\frac{\mu}{\mu+\lambda p+\lambda^* p^*}\right)^R\right] \tag{44}$$

and

$$P(D)=0. \tag{45}$$

NUMERICAL ILLUSTRATIONS

In Figure 2, graphs have been drawn to illustrate numerically the result in (43). For assumed values of λp, $\lambda^* p^*$ $P(A)$ is drawn against R for different values of μ. From the graph we observe the following:

(a) We find that after a certain number of weapons, the further increase in the number of weapons does not appreciably affect the win probability of duelist A. Thus for a given failure rate there is an optimum value of R, which ensures $P(A)$ to be nearly as high as for the case when R is infinite or $\mu=0$. Let this value of R be denoted by $\hat{R}$, then for $\mu=2$, $\hat{R}=4$; $\mu=5$, $\hat{R}=6$; and $\mu=10$, $\hat{R}=8$.

(b) When the kill rates λp and $\lambda^* p^*$ are equal, the asymptotic convergence of $P(A)$ to the value $\lambda p/(\lambda p+\lambda^* p^*)$ (win probability for A in the unlimited case) is slower, compared to the case where $\lambda p\neq\lambda^* p^*$.

(c) As the kill rate increases, the $P(A)$ increases and the effect of R becomes less significant.

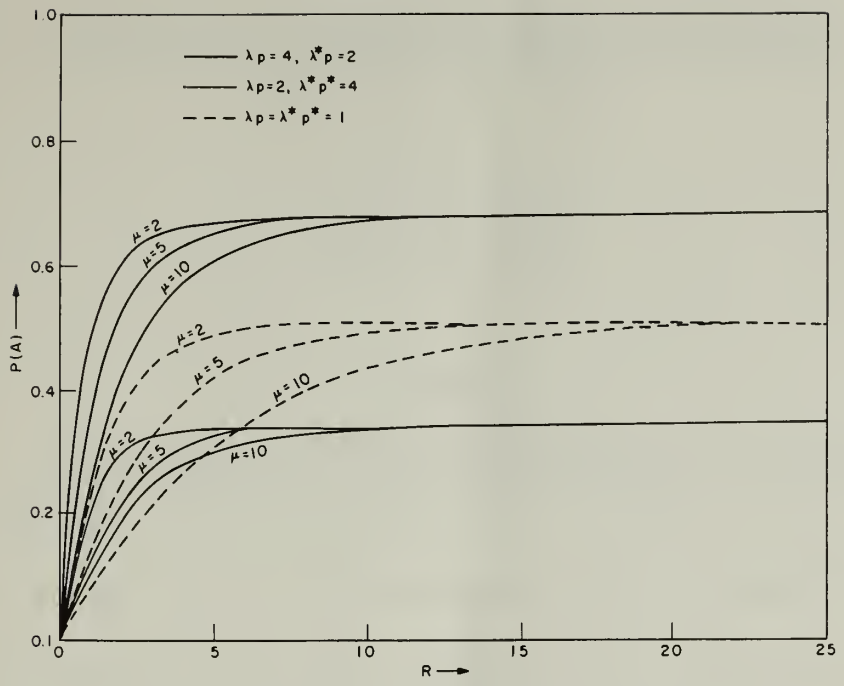


FIGURE 2. Effect of limited weapons on win probability

ACKNOWLEDGEMENTS

Thanks are due to Dr. Kartar Singh, Director, Defence Science Laboratory, Delhi-6, for permission to publish this paper. The author is extremely grateful to Prof. R. S. Varma, Dean, Faculty of Mathematics, Delhi University, and Dr. N. K. Jaiswal, Head, Statistics and Operational Research Division, Defence Science Laboratory for actively guiding him throughout the preparation of this paper.

REFERENCES

[1] Ancker, C. J., Jr., "Stochastic Duels with Limited Ammunition Supply," Operations Research *12*, 38–50 (1964).

[2] Ancker, C. J., Jr., "Stochastic Duels of Limited Time Duration," J. Canad. Operations Research Soc. *4*, 69–81 (1966).

[3] Ancker, C. J., Jr., "The Status of Developments in the Theory of Stochastic Duels," Operations Research *15*, 388–406 (1967).

[4] Pearson, K., *Tables of the Incomplete Beta-Function* (University Press, Cambridge, 1956).

[5] Williams, Trevor and C. J. Ancker, Jr., "Stochastic Duels," Operations Research *11*, 803–817 (1963).

* * *

SUBOPTIMAL ORDERING POLICIES UNDER THE FULL COST CRITERION*

Stratton C. Jaquette

Decision Studies Group
Palo Alto, California 94304

ABSTRACT

This note extends some previous work in dependent inventory models previously studied by the author. A comparison is given between optimal policies under the shortage probability criterion and a class of suboptimal policies using the usual cost minimization criterion.

This note extends the work of Iglehart and Jaquette [1] along some of the lines followed in their sequel [2].

Instead of using the shortage probability criterion, as in [2], we can impose the standard cost structure on the inventory model and ask for policies that minimized the total expected cost over the period of interest. We shall impose a shortage and holding cost on the same model discussed in [2] and present suboptimal ordering policies. This model specifies a one period delivery lag, and demands which depend on the stock on hand to start the period.

We impose a shortage and holding cost on the model which is piecewise linear and given by

$$g(u, v) = p_0 \cdot (a - v)^+ + \alpha h_0 \cdot (u - a)^+, \tag{1}$$

where u and v are the stock on hand at the beginning and end of the period, p_0 and h_0 are the shortage penalty and holding cost per item, α is the discount factor, and a is the desired level of stock which is just adequate. This level is related to the shortage level of [2] by $A = a - 1$.

If we assume that demand has a known binomial distribution and that the costs of the system are a proportional ordering cost and a shortage and holding cost given by (1), we can use the results of Iglehart and Jaquette [1] to obtain suboptimal policies. These suboptimal policies minimize the expected cost per period among the single critical number ordering policies when the resulting Markov process is assumed to be in its stationary distribution. If the unit purchase price, c, is less than the unit penalty charge, p_0, the optimal single critical number policy is to order up to the level $\bar{y}$, where $\bar{y}$ is the greatest integer, y, satisfying

$$\frac{\alpha h_0 + c\dfrac{2p}{1-p}}{\alpha h_0 + p_0} \leq \sum_{i=0}^{a-1} \binom{y}{i} \left(\frac{1-p}{1+p}\right)^i \left(\frac{2p}{1+p}\right)^{y-i},$$

where p is the parameter in the binomial distribution of demand.

The summation on the right hand side of the above inequality is just the stationary probability of being short at the end of the period under the single critical number policy, y. Had we restricted our

*Prepared under contract Nonr 4457(00) (NR-047-093) for the Office of Naval Research.

consideration of policies under the shortage probability criterion to single critical number policies and stationary analysis, we would have obtained almost the same result. If we let

$$\gamma=\frac{\alpha h_0+c\,\dfrac{2p}{1-p}}{\alpha h_0+p_0},$$

then the suboptimal policies obtained under the two criteria would be to order up to fixed levels differing by one unit, and we would order less under the cost criterion.

We can also find the optimal myopic ordering policy under the cost criterion. The myopic policy was optimal for the shortage probability criterion, but it is not generally optimal under the cost criterion.

The myopic policy is to order up to the level $\bar{y}(x)$ where $\bar{y}(x)$ is the value of y which minimizes the expected discounted cost.

$$c\cdot(y-x)+E[g(y-D_1,\ y-D_1-D_2)], \tag{2}$$

where D_1 and D_2 are the demands during this period and next period. If $\bar{y}(x)$ minimizes (2) then the policy of ordering up to $\bar{y}(x)\vee x$ is the optimal myopic policy, and the optimal myopic cost is given by

$$c(x)=\min_{y\geq x}(y-x)+\alpha E[g(y-D_1,\ y-D_1-D_2)].$$

For a known binomial demand with parameter p, $\bar{y}(x)$ is the value of y which minimizes

$$cy-cx+\alpha p_0\sum_{k=0}^{a-1}(a-k)\binom{x}{y-k}p^{y-k}(1-p)^{x-y+k}$$

$$+\alpha^2 h_0\sum_{k=a}^{\infty}(k-a)\sum_{i=0}^{x}\binom{x}{i}\binom{y-i}{k}p^{y-k}(1-p)^{x-i+k}.$$

If we assume that the binomial parameter is unknown, and we use the Bayesian estimation procedures of [2], then the optimal mypoic policy is to order up to $\bar{y}(x)$ where $\bar{y}(x)$ minimizes

$$cy-cx+\alpha p_0\sum_{k=0}^{a-1}(a-k)\binom{x}{y-k}(\beta+1)\binom{\beta}{\alpha}/(\beta+x+1)\binom{\beta+x}{\alpha+y-k}$$

$$+\alpha^2 h_0\sum_{k=a}^{\infty}(k-a)\sum_{i=0}^{x}\binom{x}{i}\binom{y-i}{k}(n+1)\binom{n}{m}/(n+x+y-i+1)\binom{n+x+y-i}{m+y-k},$$

where we use an *a priori* distribution for the parameter p which is a beta distribution with parameters n and m.

REFERENCES

[1] Iglehart, D. L. and S. C. Jaquette, "Multiclass Inventory Models with Demand a Function Of Inventory Level," Naval Research Logistics Quarterly *16*, pps. 495–502 (1969).

[2] Iglehart, D. L. and S. C. Jaquette, "Optimal Policies under the Shortage Probability Criterion for an Inventory Model with Unknown Dependent Demand," Naval Research Logistics Quarterly *16*, pps. 485–493 (1969).

* * *

OPERATIONAL DEFINITIONS OF INVENTORY RECORD ACCURACY*

D. A. Schrady

Department of Operations Analysis
Naval Postgraduate School,
Monterey, California

ABSTRACT

An inventory stock record is in error when the information on the stock record is not in agreement with the actual physical situation. We address the questions of what is meant by inventory record accuracy as reported in the literature and what should be meant by this term, in the context of the inventory record accuracy problem defined by the Naval Supply Systems Command. The need for, and suggestions of, operational definitions of error measures are demonstrated in terms of the reporting of accuracy statistics, the formulation of inventory record accuracy goals, and the determination of corrective measures.

1. INTRODUCTION

The problem concerns errors which are generated in and exist in inventory stock records. In a broad sense, the problem embraces the subject areas of accounting and audit systems and controls, personnel, physical security, work methods and standards, management information systems, and mathematical inventory theory.

This report is limited to the discussion of what is meant by the stock record error, as a first step toward the quantitative analysis of the costs and benefits associated with record accuracy. Specifically, we address the inventory record accuracy problem as it exists in the Navy. The Auditor General of the Navy Report of April 1965, indicated that 29 percent of the Navy and Marine Corps stock records were in error [1]. This error is a source of embarrassment to the Supply Corps, opens them to criticism from the General Accounting Office [4], compromises their budget requests, and, most significantly, reduces their ability to perform their mission of supplying the operating units of the service.

Inventory record accuracy (the lack of) is a problem confronting all of the military services, and probably industry as well. It is an old problem and one not easily remedied. It is an especially difficult problem for the military establishment due to the range of items stocked. Only the very largest of corporations have as many as 100,000 different items in inventory. By contrast, the Naval Aviation Supply Office (one of three inventory managers in the Navy supply system) alone carries more than 400,000 items and a large Naval Supply Center stocks as many as 900,000 items. The sheer volume of record keeping aggravates the inventory record accuracy problem. This volume also dictated early use of electronic data processing (EDP) systems. While it is always fashionable to state that "people are no damn good" (and indeed the statement applies to certain aspects of this problem), EDP has created a physical gap between the people and the records and has eroded the structure of responsibility. This complicates the problem of error correction.

Some inventory record errors are the result of theft of material. The cost of errors generated by thefts is the cost of the material involved and the loss of service that the material could have provided;

*This task was supported by the Research and Development Division, Naval Supply Systems Command, under NAVSUP RDT&E task area No. 38.511.003.

however, the majority of record errors are generated by "honest mistakes." The cost of these honest mistakes has a monetary value in terms of increased investment levels (when this results), but is best measured in terms of *disservice*. A recent Navy Area Audit Service report [5] went so far as to state, ". . . We find even less assurance that the records are sufficiently reliable to permit even adequate material support to operating forces."

Thus, the inventory record accuracy problem is real and serious, and it affects the ability of the Supply Corps to serve the operating forces.

The motivation behind this study is that, in spite of the reports and studies already conducted on inventory record accuracy, there does not appear to be a clear understanding or definition of the terms and measures of inventory record accuracy and/or error. Until we define our terms and have some common basis on which to communicate and take measurements, the study, reporting, and formulation of corrective actions for record errors will remain unnecessarily vague and imprecise. The statement that 29 percent of the Navy and Marine Corps inventory records are in error is not necessarily meaningful. One must know, at the very least, how the figure was obtained, what constituted an error, and the general nature of the individual discrepancies.

By way of introduction, we note that an *error* is a discrepancy between what the stock record indicates as the "situation" and what is actually the "situation." Many types of discrepancies are possible. The most widely recognized error is a discrepancy in the on-hand quantity.

The mechanism by which these discrepancies are discovered and reconciled is the physical inventory. "Physical inventory" connotes a program to count the quantity of an item in storage, to compare with that recorded on the accountable stock record and to adjust the record where discrepancies exist. Several types of physical inventories are possible. One type of inventory is the "wall-to-wall" inventory in which the on-hand quantity of every item is verified at a specified point in time. Alternatively there are "spot" inventories which are conducted on specified items at any time that the situation warrants. Commonly, a spot inventory is requested for an item if the records indicate a zero balance or if there has been a warehouse refusal.

NAVSUP policy with respect to the taking of wall-to-wall inventories has changed several times during the past 10 years. Declining funding dictated a move from an annual physical inventory to a tri-annual schedule. The tri-annual schedule was deemed unsatisfactory and a plan of conducting annual inventories on "active" items only was adopted. Most recently, a spot-inventory-prior-to-procurement scheme has been instituted. The Fleet Material Support Office is currently developing a "trigger" scheme where the triggering of a spot inventory is based on item demand characteristics.

The changes have been made in an effort to be more responsive to the problem of record accuracy within a framework of limited resources. Two factors combine to reduce the perceived effectiveness of physical inventory programs. The lack of measurement of the results of expenditures on physical inventories has resulted in decreased funding. In this climate of insufficient funds and personnel, the physical inventory process has been neglected in favor of functions directly related to supply operations.

2. LITERATURE

If we create a distinction between reports of measurements of inventory record accuracy and analyses of the inventory record accuracy problem, then we can state that the literature contains few reports of analyses of the problem. The General Accounting Office, Navy Auditor General, and Navy Area Audit Service reports generally only establish and document the existence of the problem. As a part of the background on inventory record accuracy we very briefly note the few analyses of the inventory record accuracy problem and their conclusions.

The first study we note was conducted for the Army Office of Ordnance Research [6]. They found that, contrary to one's intuition, the paper work of issue and receipt processing contributes only about 20 percent of the total record error and the remaining 80 percent is contributed by the error correcting procedures themselves; i.e., the physical inventory process. They also define a *residual error*, the error remaining in the records after completion of the physical inventory, and determined it to be on the order of 20 to 25 percent. The second study noted here was conducted by the Naval Supply Research and Development Facility at Bayonne (NAVSUPRANDFAC) [3]. They found a residual error of 7 percent, but there is reason to suspect that this figure is optimistic. They also determined relationships between the probability of an incorrect count and the on-hand quantity, the growth of record errors as a function of the time since the last physical inventory, and the probability of a record being in error at the end of some time interval as a function of the demand for the item during the period. Finally, the recent report of the Control of Inventory Task Group [2] contains some quantitative data, among which is a study showing that 55 percent of the error was introduced by the physical inventory process.

The inventory record accuracy problem is about, and punctuated with, error percentages. The common term in the literature is "error rate," but it is legitimate to question the meaning of this term. For that matter, how is or should an error be defined? These are not trivial questions and will be addressed in the next section.

3. THE DIMENSIONS OF STOCK RECORD ERROR

Let us begin by indicating what was being measured in two of the analyses referenced in the previous section. In the NAVSUPRANDFAC report, the only *discrepancies* noted as *errors* were discrepancies in the on-hand quantity. If the item could not be found or could not be identified, then it could not be counted. Location and identity discrepancies were considered only as they contribute to on-hand quantity discrepancies. The "error rates" cited should be interpreted in the following way: At a given point in time, in a sample of N items, there were n item records in which there was discrepancy between the recorded amount on hand and the actual warehoused amount on hand, and therefore $100(n/N)$ percent of the records were in error. Note that *all* discrepancies in the on-hand quantity were counted equally as errors. This treatment ignored the absolute magnitude and sign of the discrepancies, the magnitude of the discrepancies as percentages of the actual on-hand quantities, and the dollar values of the discrepancies.

While the definition of error was never explicitly given in the NAVSUPRANDFAC report [3], the Army report [6] used the following definition of error: "A difference between record balance and physical stock is counted as a discrepancy (error) if the difference is 1% or more of the record balance, or if the monetary value of the difference is $1 or more."
The above statement is hereafter referred to as a criterion for distinguishing between major and minor errors.

Toward developing the concept and dimensions of error, we shall question, *for purposes of illustration only*, the meaning of the recommended NAVSUP Specific Objective for inventory record accuracy [7]: "A program for achieving and maintaining a minimum 90 percent item inventory record accuracy." The 90 percent requirement is subject to several interpretations, even if we defer the question of what constitutes an accurate item record. It is helpful to recognize the two aspects of the statement of the Specific Objective: the *performance* implied in the 90 percent figure and the *definition* of record accuracy. Each of these aspects of inventory record accuracy are discussed in the sections which follow.

4. INVENTORY RECORD ACCURACY PERFORMANCE

Several possible interpretations of the performance measure are as follows:

I. At any arbitrarily selected time at least 90 percent of the records are accurate;

II. The time average of the percent of accurate records is at least 90 percent;

III. The time average of the accuracy of an (every) individual item record is at least 90 percent.

These interpretations of the 90-percent figure can be explained in terms of an example population of 10 item records, designated item records A, B, . . ., J. Let the times at which accuracy checks are made be denoted as t_1, t_2, ... For example, consider the time between successive checks to be a month; then 12 checks constitute a year. Given that we have some operational definition of record accuracy, at each time a check is made each item record may be classified as accurate or in error. These results will be displayed in a matrix with an accurate record indicated by a + and an inaccurate record indicated by a 0. Figure 1 exhibits four such matrices showing the condition of each item record at the time of each of 12 accuracy checks.

Given the first interpretation of 90-percent record accuracy, we require that *in any and all monthly accuracy checks 9 or 10 of the 10 records be accurate.* The matrices of Figure 1, (a) and (b), represent acceptable performance under this interpretation of performance, while the matrices of Figure 1, (c) and (d), represent unacceptable performance. Implicit in the above statement is a notion of the total *time period* over which the performance criterion is applied. The implied time period is best taken to be infinite; i.e., over all future time. Thus, Figure 1(c) fails because all item records are in error in the eighth test, and Figure 1(d) fails because two records are in error in the tenth test. While Figure 1, (a) and (b), represent acceptable performance, the difference, of course, is that in Figure 1(a) the record of item A is permanently in error.

The second interpretation of 90-percent record accuracy relaxes the requirement that record accuracy be 90 percent or greater at each check and requires only that the time average be 90 percent or greater. Let n be the total number of records (10 in this example), n_i be the number of accurate records in the ith check $(0 \leq n_i \leq n)$, and N be the number of checks over which the time average is to be computed. Here, N is equal to 12, and the time period over which the performance criterion is to

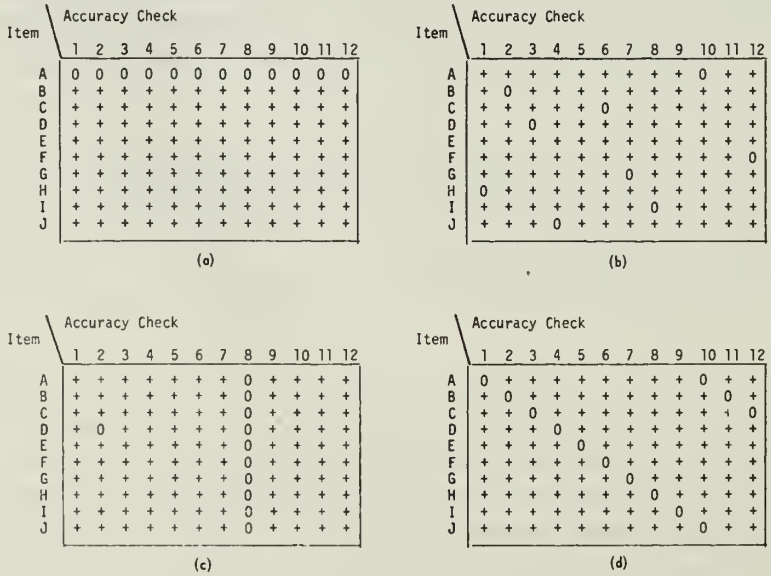

Item
Accuracy Check
1 2 3 4 5 6 7 8 9 10 11 12
A 0 0 0 0 0 0 0 0 0 0 0 0
B + + + + + + + + + + + +
C + + + + + + + + + + + +
D + + + + + + + + + + + +
E + + + + + + + + + + + +
F + + + + + + + + + + + +
G + + + + + + + + + + + +
H + + + + + + + + + + + +
I + + + + + + + + + + + +
J + + + + + + + + + + + +
(a)
Item
Accuracy Check
1 2 3 4 5 6 7 8 9 10 11 12
A + + + + + + + + + 0 + +
B + 0 + + + + + + + + + +
C + + + + 0 + + + + + + +
D + + 0 + + + + + + + + +
E + + + + + + + + + + + +
F + + + + + + + + + + + 0
G + + + + + + 0 + + + + +
H 0 + + + + + + + + + + +
I + + + + + + + 0 + + + +
J + + + 0 + + + + + + + +
(b)
Item
Accuracy Check
1 2 3 4 5 6 7 8 9 10 11 12
A + + + + + + + 0 + + + +
B + + + + + + + 0 + + + +
C + + + + + + + 0 + + + +
D + 0 + + + + + 0 + + + +
E + + + + + + + 0 + + + +
F + + + + + + + 0 + + + +
G + + + + + + + 0 + + + +
H + + + + + + + 0 + + + +
I + + + + + + + 0 + + + +
J + + + + + + + 0 + + + +
(c)
Item
Accuracy Check
1 2 3 4 5 6 7 8 9 10 11 12
A 0 + + + + + + + + 0 + +
B + 0 + + + + + + + + 0 +
C + + 0 + + + + + + + + 0
D + + + 0 + + + + + + + +
E + + + + 0 + + + + + + +
F + + + + + 0 + + + + + +
G + + + + + + 0 + + + + +
H + + + + + + + 0 + + + +
I + + + + + + + + 0 + + +
J + + + + + + + + + 0 + +
(d)

FIGURE 1. Record accuracy matrices

be applied is explicit. Then the population accuracy in the ith check is n_i/n, and the time average of the population record accuracy over N time periods is

$$\text{Time Average} = \frac{1}{nN} \sum n_i.$$

Thus any matrix with 12 or less zeros (records in error) is acceptable under this interpretation (12/120 = 10 percent). Under this less restrictive interpretation of performance, Figure 1 (a), (b), and (c) all represent acceptable performance. Only Figure 1(d) represents unacceptable performance.

The third interpretation of the 90 percent accuracy requirement specifies that the time average of the accuracy of *each and every* item record be 90 percent or greater. This interpretation prohibits any record from remaining permanently in error (as in Figure 1(a). In our example, *interpretation III requires that any row (record accuracy history) have at most one zero.* Figure 1(b) represents acceptable performance as the time average accuracy of all the item records is 92.5 percent or greater. Figure 1(c) represents unacceptable performance because the time average accuracy of item record "D" is only 83.3 percent. Similarly, Figures 1(a) and (d) represent unacceptable performance under this interpretation.

In the table below are summarized the acceptability (A) or unacceptability (U) of each of the performance matrices with respect to each of the interpretations.

Interpretation	Matrices			
	1a	1b	1c	1d
I	A	A	U	U
II	A	A	A	U
III	U	A	U	U

Note that the performance represented in Figure 1(d) was not acceptable under any of the three interpretations. However, at least intuitively, the performance in Figure 1(d) is preferable to the performances of Figures 1(a) and (c). Figures 1(a) and (c) represent extreme cases and may not fairly test the adequacy of the proposed performance interpretations. However, the point raised, whether or not 1(d) is preferable to either or both of 1(a) and (c), is indicative of pitfalls that one may encounter in designing and applying performance indices.

Finally, we note the critical influence of the time parameter on the two interpretations which involve time averages. We had been thinking of the time intervals at which accuracy checks were made as months. The time averages of interpretations II and III were computed over a time period of 1 year. Suppose now that we are still interested in time averages computed over 1 year, but that we now assume that the data in the four tables represents accuracy checks for 3 years at quarterly intervals. (For example, DOD regulations require quarterly physical inventories for high-value-asset-control material.) While the matrix of Figure 1(b) was *acceptable* under interpretation III with monthly checks,

it is *unacceptable* under interpretation III in each of the 3 years now represented by the 12 quarterly checks. The following table summarizes the 3 years of quarterly accuracy checks represented in Figures 1(a)–1(d) in terms of the two performance interpretations involving time averages.

Interpretation	Figure 1			
	(a)	(b)	(c)	(d)
	1 2 3	1 2 3	1 2 3	1 2 3
II	A A A	A A A	A U A	A A U
III	U U U	U U U	U U A	U U U

The above discussion was intended to call attention to the need to define operationally any statement or objective with respect to inventory accuracy performance. In discussing the meaning of record accuracy and/or error, we have tried to separate the question of performance from the question of error. We now address the meaning of a record error.

5. THE DEFINITION OF INVENTORY RECORD ERROR

A stock record for a given item contains a large amount of information. Any of the bits of data about the item could be discrepant. The question is one of deciding which discrepancies constitute errors and how error rates or error percentages should be defined. A record could be in error with respect to the on-hand quantity, primary storage location, identity (cog, FSC, FIIN), price, unit of issue, quantity on order, substitutes, etc. Are discrepancies in all of these bits of information equally important? Does the question of error have to be an all or nothing characteristic; i.e., any discrepancy means the record is in error and only discrepancy-free records are accurate. Is the magnitude of discrepancy important? Is a quantity error of 10 units, plus or minus, as important when the actual on-hand quantity is 869 units as when the actual on-hand balance is 4 units? Regardless of the actual quantity on hand is the importance of a quantity discrepancy of 10 units independent of the unit price of the item? Should there be some concept of major and minor errors, or of the degree or seriousness of a given discrepancy?

The Army report [6] used an explicit definition of major and minor errors (cited earlier in this paper). Most would agree that an error representing less than 1 percent of the on-hand quantity *and* less than 1 dollar in value is not worth worrying about. The question, of course, arises as to whether an error representing 3 percent of the on-hand quantity and $4.53 in value is serious. An all-encompassing answer to this type of question is not easily obtained.

Another point we wish to raise is that much of the data on an item stock record is not normally verified. While the on-hand quantity and location are verified by periodic physical inventories, there are no system-wide programs to verify such data as unit price, unit of issue, on order quantity, etc. But errors in "unit of issue" result with certainty in on-hand quantity errors being introduced. Further, at the Inventory Control Point, procurement decisions are tied to the *inventory position* of the item. The inventory position is defined as the quantity on hand, plus the quantity on order, minus the quantity backordered. The on-hand quantity is but one of three pieces of information required to implement

current UICP inventory policies. In today's climate of low procurement budgets, the on-hand quantity may not even be the most significant of the three quantities which determine the inventory position. Thus the "buy too soon" and "buy too late" costs associated with inaccurate inventory records may be primarily dependent upon discrepancies which are not now being observed and corrected, even periodically.

We have raised a number of questions and, of course, it is easier to raise such questions than it is to answer them. However, in the next section, we propose and evaluate a number of ways of dealing with on-hand quantity discrepancies.

6. POSSIBLE MEASURES OF INVENTORY RECORD ERROR

In this section we investigate a number of inventory record error measures, all based on discrepancies in on-hand quantities. Various measures are computed using, for purposes of exposition, a reconstruction of the data of "Sample I" of the NAVSUPRANDFAC Report [3]. This sample contains identity, recorded on-hand quantity, actual on-hand quantity, price, and demand information for 508 items. Ninety nine of these item records show on-hand quantity discrepancies.

The *raw error percentage*, defined here as the percent of item records with any on-hand quantity discrepancy, for this sample was 19.5 percent. This raw error percentage is a legitimate measure of record error, but it is by no means the only legitimate or meaningful error measure. It ignores the degree of seriousness of the individual discrepancies, the financial gains and losses involved, and the impact of these discrepancies on supply effectiveness.

If the major error criterion of the Army report [6] was applied to the data, six errors each representing on-hand quantity discrepancies of less than 1 percent and less than \$1 in value would be deleted (not counted as errors). The *major error percentage* would thus be $100(93/508) = 18.3$ percent. If major error was defined as one exceeding 5 percent of the on-hand quantity and representing more than \$5 in value, then 27 of 99 discrepancies would not be counted as errors and the error percentage would be $100(72/508) = 14.2$ percent.

A more elaborate scheme might call for assessing the seriousness of each on-hand quantity discrepancy in terms of the actual quantity on hand. Such a scheme recognizes the difference between a discrepancy of 10 units when only 5 units are actually on hand versus the same discrepancy when 500 units are actually on hand. Let us define, for the ith item, a_i and r_i as the actual on-hand quantity and recorded on-hand quantity, respectively. We then define

$$\Delta_i = |a_i - r_i|,$$

and

$$m_i = \max\ (a_i, r_i)\ .$$

The error measure for individual items is then Δ_i/m_i, $0 \leq \Delta_i/m_i \leq 1$, $m_i > 0$. If $a_i = r_i$, then $\Delta_i = 0$ and no contribution is made to the error measure. If the total error for a population of N items is defined as

$$1/N \sum_{i=1}^{N} \Delta_i/m_i,$$

then the total on-hand weighted error will also lie in the unit interval and

$$100\left[\frac{1}{N}\sum \Delta_i/m_i\right]$$

is the *on-hand weighted error percentage.*

Applied to the sample data, the on-hand weighted error percentage is 7.8 percent. An item with an on-hand quantity discrepancy of 890 units, but actual on-hand quantity of 7000 units, had an error measure computed to be $890/7000 = 0.11$. While the discrepancy was large in absolute terms, it was small in a relative sense. We have, of course, not considered the financial implication of the discrepancy. Another difficulty or shortcoming of the on-hand weighted error measure is its inability to discriminate between serious errors. Within the sample are three items which, for present purposes, we label items 1, 2, and 3. Item 1 had a recorded on-hand quantity of 0 units and an actual on-hand quantity of 800 units. The actual and recorded amounts for the other two items were: $a_2 = 1$, $r_2 = 0$; $a_3 = 0$, and $r_3 = 1$. The on-hand weighted error measures for all three of these is 1.0, the maximum possible error, though intuitively these quantity discrepancies are not all equally significant.

The financial aspects of on-hand quantity discrepancies can be seen in costing the individual discrepancies. The records indicated a total value for the sample inventory of 508 items of $67,648. The actual total investment was $70,762. If the unit cost of the ith item is denoted as c_i, and if a_i, r_i, and Δ_i are as defined previously, then we compute the *net dollar error percentage* as

$$100 \frac{\Sigma c_i(a_i - r_i)}{\Sigma c_i a_i}.$$

For the sample inventory data, the net dollar error percentage was $3114/$70762 = 4.4 percent.

If we consider the total or gross dollar error in investment, however, the error percentage is much higher. In the net dollar error percentage an overstatement of investment on one item may be canceled out by an understatement of investment on another item. Adding all investment discrepancies, regardless of their sign, results in a *gross dollar error percentage* defined as

$$100 \frac{\Sigma c_i \Delta_i}{\Sigma c_i a_i}.$$

For the sample, the gross dollar error percentage was $13687/$70762 = 19.3 percent.

Another way to look at investment discrepancies is by means of a breakdown of the discrepancies by dollar amount intervals. For the sample the breakdown of the 99 discrepancies is as follows:

$ Interval	*Number of Stock Records*
0 to 1.00	32
1.01 to 10.00	30
10.01 to 100.00	22
100.01 to 1000.00	9
1000.01 to 3000.00	6

One third of the discrepancies had a value of $1 or less, and nearly two thirds of the discrepancies had a value of $10 or less. A relatively small percentage of the items contributed most of the total investment discrepancy.

Finally, we address inventory record error measures which are directly related to short-term supply effectiveness. On-hand quantity errors may have a direct short-term influence on effectiveness if the discrepancies can give rise to a warehouse refusal or an unnecessary referral. A warehouse refusal

may be generated anytime there exists an $r_i > 0$, $a_i = 0$ situation; i.e., the records indicate that stock is on hand, but, in fact, the on-hand balance is zero. An unnecessary referral may be generated by a $r_i = 0$, $a_i > 0$ situation.* In the sample, there were at least 9 discrepancies which could generate warehouse refusals and at least 16 discrepancies which could have generated needless referrals. Thus, 25 of the 508 item records had discrepancies which could have lead directly to a degradation of short-term supply effectiveness. What we might call *the short-term effectiveness degradation measure*, a percentage, would then be $25/508 = 4.9$ percent.

We have shown that, for the sample inventory record data, error measures of between 4.4 and 19.5 percent may be computed. One could choose his measure to fit his needs, but the question of inventory record accuracy is not simply a numbers game. One conclusion that might be drawn from this section is that no single error measure is entirely satisfactory. At a minimum, one would probably want to know the inventory error situation in terms of the percentage of major errors, the net investment discrepancy, and the outlook for short-term supply effectiveness.

7. CONCLUSION

We have discussed the meaning of stock record error and inventory record accuracy as a first step toward the quantitative analysis of the costs and benefits associated with record accuracy. In terms of inventory operations, stock record errors present two basic problems. The first concerns the application of optimal ordering policies to inaccurate records. If the records indicate more stock than is actually present, the system fails to order when it should with the result that greater than optimal shortages are incurred. If the records indicate less stock than is actually present, the system reorders too soon with the result that the inventory investment is too large. The second problem is one of minimizing total costs. There is a cost associated with operating a system with inaccurate inventory records, and there is a cost associated with achieving and maintaining a given level of inventory record accuracy. The problem is to specify the level of record accuracy which minimizes the sum of these costs. A subproblem is to specify the procedures for obtaining a given level of record accuracy at minimum cost.

Finally, we note that the problem of record accuracy exists wherever records are kept. For example, another area in which record accuracy is known to be a problem is in personnel management.

REFERENCES

[1] Auditor General of the Navy, "Accuracy of Navy and Marine Corps Stock Records," Service-Wide Audit Report No. IA1–65 (Apr. 26, 1965).

[2] Control of Inventory Task Group, "Progressive Onslaught to Stamp Out Stock Errors," final report to the Naval Supply Systems Command (Nov. 1967).

[3] Emma, Charles King, "Physical Inventory Procedures at Navy Stock Points—Observations on Physical Inventory and Stock Record Accuracy," Naval Supply Research and Development Facility, Bayonne, N.J. (June 1966) (Available from DDC, AD 488–317L).

[4] General Accounting Office, "Improved Inventory Controls Needed for the Departments of the Army, Navy, and Air Force and the Defense Supply Agency," Report to Congress (Nov. 1967).

*The discussion of warehouse refusals and referrals should consider the requisition quantity. A warehouse refusal could be generated by an $r_i = 8$, $a_i = 4$ situation if the requisition was for 5 units of the item. The ($r_i > 0$, $a_i = 0$) and ($r_i = 0$, $a_i > 0$) conditions used above put only a lower bound on the loss of effectiveness due to error caused refusals and referrals.

[5] Navy Area Audit Service Norfolk, "Regional Audit of Internal Controls Relating to Inventory Management in the Fifth and Sixth Naval District" (June 19, 1967).

[6] Rinehart, R. F., "Physical Inventory Accounting Program/Technical Report Number 3," Office of Ordnance Research, U.S. Army, Durham (Mar. 1960) (Available from DDC, AD237078).

[7] Seymour, John, "Inventory Record Accuracy," A Naval Supply Systems Command Briefing to the Executive Board (July 24, 1967).

* * *

A NOTE ON THE CORE OF A COOPERATIVE GAME WITHOUT SIDE PAYMENT†

Yasoi Yasuda

Tokyo Institute of Technology
Tokyo, Japan

The purpose of the present note is to establish a theorem of the core of a cooperative game in Aumann [2] in the form of changing the assumption, and moreover to reexamine the axioms in light of the core of a game and to suggest a small and reasonable change (in the axioms) that has been suggested by Refs. [1–5].

In this note we don't need the knowledge of the theory of a cooperative game without side payment by Aumann and Peleg [2] and [3].

1. Aumann and Peleg suggest in [2] and [3] the definition of a game in characteristic function form. Definition: A characteristic function is a pair (N, v), where N is a finite set, and v is a function that carries each subset S of N into a subset of E^N so that

(1) $v(S)$ is convex;

(2) $v(S)$ is closed;

(3) $v(\phi)=E^N$;

(4) $v(S)$ is a comprehensive, i.e., if $x\in v(S)$ and $y^S \leqq x^S$, then $y\in v(S)$;

(5) if $S\cap T=\phi$, then $v(S\cup T)\supset v(S)\cap v(T)$; A game in characteristic function form, or simply a game, is a triple (N, v, H), where (N, v) is a characteristic function, and

(6) H is a convex compact polyhedral subset of E^N; We shall say that (N, v, H) is an ordinary game if

(7) $x \in v(N)$ if and only if there is a $y \in H$ such that $x \leqq y$. We shall need the basic difinitions as follows:

A payoff vector x is said to dominate a payoff vector y via S(notation; $x \succ_S y$) if $x \in v(S)$ and $x^S > y^S$;

x is said to dominate y(notation; $x \succ y$) if there is an S such that $x \succ_S y$;

For $x, y \in E^N$ and $S \subset N$, Dom $_S x=\{y: x \succ_S y\}$;

For $x, y \in E^N$, Dom $x=\{y: x \succ y\}$;

For $K \subset E^N$ and $S \subset N$, Dom $K_S=\cup_{x\in K}$ Dom $_S x$;

For $K\subset E^N$, Dom $K=\cup_{x\in K}$ Dom x;

V is K stable if $V=K-$Dom V;

The K-core (notation; $C(K)$) is the set K-Dom K;

We denote $v^i=\sup_{x\in v(\{i\})} x^i$;

x is individually rational if $x^i \geqq v^i$ for all $i\in N$;

x; x is group rational if there is no $y\in H$, such that $y > x$;

We denote†† $H_i=\{x;\ x\in H,\ x$ is individual rational$\}$,

$Hg=\{x;\ x \in H,\ x$ is group rational$\}$, and

$$Hig=Hi \cap Hg.$$

---

†I am indebted to Professor M. Suzuki and especially to Mr. Y. Ito in this note.

††In Aumann [2], denoted: $E=H_g$, $A=H_i$, $A=H_i$, $A=H_{ig}$, and $E=H$.

2. Aumann and Peleg, in [1–3], say that the following core theorem is true and true only because H is a polyhedron.

THEOREM†: If (N, v, H) is an ordinary game, then

$$C(H)=C(H_i)=C(H_g)=C(H_{ig})$$

and Aumann [2] gives a counter-example if H is not polyhedral.

A counter-example by Aumann [2]:

let $N=\{i, 2, 3\}$ and H be the convex hull of the sets C and D, where

$$C=\{x : x^1 \geq 0,\ x^2 \geq 0,\ x^3=0,\ (x^1)^2+(x^2)^2 \leq 1\},$$
$$D=\{x : x^1 \geq 0,\ x^2 \geq 0,\ x^3=1,\ (x^1)^2+(x^2+1)^2 \leq 4\}.$$

Then

$$Hg=D \cup \{x;\ x^1=0,\ x^2=1,\ 0 \leq x^3 \leq 1\}.$$

Define the characteristic function v by

$$v^N=0,\ v(\{ij\})=\{x;\ x^i \leq 1/2 \text{ and } x^j \leq 1/2\},$$

$$v(N)=\{x;\ \text{there is a } y \in H \text{ such that } y \geq x\}.$$

we have $(1/2, 1/2, 1/2) \in H$ and $(1/2, 1/2, 1/2) \leftarrow \{13\}\ (0, 1, 0)$, but there is no member of H_g that dominates $(0, 1, 0)$. That is, $C(H)$ does not coincide with $C(H_g)$.

The theorem is true if H is a polyhedron, and a counter-example is true, too. But we can prove the theorem in the case where H is not a polyhedron, and avoid this counter-example as shown by the following:

3. We must establish the core theorem in the case where H is not a polyhedron.

For any convex compact subset H of E^N we define the set H^* that is constructed from H.

$H^*=H_+ \cap H_-$, where
$H_+=\cup_{x \in H}\ H_+(x)$
$H_-=\cup_{x \in H}\ H_-(x)$
$H_+(x)=\{y;\ y \in E^N, \text{ for a } x \in H \text{ such that } y \leq x\}$
$H_-(x)=\{y;\ y \in E^N, \text{ for a } x \in H \text{ such that } y \geq x\}$,

then H^* is the convex compact subset of E^N, and $H \subset H^*$. Moreover, clearly $(H^*)^*=H^*$. H^* is described as the union of the following disjoint sets:

$$H^*=(H_+^o \cap H_-^o) \cup (H_+^o \cap H_-^f) \cup (H_+^f \cap H_-^o) \cup (H_+^f \cap H_-^f),$$
(direct sum)

†This theorem corresponds to Theorem 6 in [1], Theorem 8 in [2], and Theorem 3 in [3].

where

$$H_+^o = \text{the interior of } H_+$$

$$H_-^o = \text{the interior of } H_-$$

$$H_+^f = \text{the frontier of } H_+$$

$$H_-^f = \text{the frontier of } H$$

Moreover, define

$$H_+^* = \cup_{x \in H^*} H_+^{*o} + (x),\ H_-^{*o} = \cup_{x \in H^*} H_-^{*o}(x),$$

where

$$H_+^{*o}(x) = \{y : y \in E^N, \text{ for a } x \in H^* \text{ such that } y < x\}$$

and

$$H_-^{*o}(x) = \{y : y \in E^N, \text{ for a } x \in H^* \text{ such that } y > x\}.$$

For $x > 0$, define $f(x) = \max i, j \in N(x^i/x^j)$.

We have

(1) $f(x) = 1$ and $f(y) = 1$ imply $f(x+y) = 1$.

We need the following lemma to verify the corresponding theorem:

LEMMA 1†: For every closed convex set h in E^N, there is a positive number 1 such that for all $x \in H^* - H_g^*$, there is an $x' \in H^*$, such that

(2) $x' > x$ and $f(x' - x) = 1$.

PROOF: $H_g^* = (H_+^f \cap H_-^o) \cup (H_+^f \cap H_-^f)$ (direct sum), so we have

$H^* - H_g^* = (H_+^o \cap H_-^o) \cup (H_+^o \cap H_-^o)$ (direct sum), therefore

$H^* - H_g^* \subset H_+^o$.

For any $x \in H^* - H_g^*$ since H_+^o is an open set, there is a positive number ϵ such that

$N_\epsilon(x) \subset H_+^o$, where

$N_\epsilon(x)$ is an ϵ neighborhood of x. Let's define

$M_\epsilon(x) = \{y; y \in N_\epsilon(x), \text{ for a } x \in H^* - H_g^* \text{ such that } y > x\}$. Then,

$$M_\epsilon(x) \subset N_\epsilon(x) \text{ and } M_\epsilon(x) \subset H_-^{*o}(x)$$

define $y_\delta = x + \delta a$. where

$$a = (1, 1, \ldots\ldots\ldots, 1)$$

for sufficiently small positive δ such that $0 < \delta < \epsilon$ $y_\delta > x$, so we have $y_\delta \in M_\epsilon(x)$.

$$M_\epsilon(x) \subset H_-^{*o}(x),\ H_-^{*o}(x) \subset H_-^{*o}, \text{ and } H_-^{*o} = H_-^o \text{ imply } y_\delta \in H_-^o.$$

Moreover, since $M_\epsilon(x) \subset N_\epsilon(x)$ and $N_\epsilon(x) \subset H_+^o$, we have $M_\epsilon(x) \subset H_+^o$, so $y_\delta \in M_\epsilon(x)$ implies $y_\delta \in H_+^o$. After all $y_\delta \in H_+^o \cap H_-^o$.

$$H_+^o \cap H_-^o = H^{*o},\ H^{*o} \subset H^*,$$

therefore, $y_\delta \in H^*$. We define $x' = y_\delta$ for this δ. Then,

$$x' - x = \delta a$$

---

†This lemma corresponds to lemma 2 in [2].

so that $f(x'-x)=f(a)$, as $f(a)=1$, we obtain

$$f(x'-x)=1.$$

This completes the proof. We shall need a norm on E^N. For $x \in E^N$, we define

$$||x||=\max_N|x^i|.$$

In addition to the usual norm properties, we obtain (3) if $x>0$ and $y>0$, then $\|x+y\|>\max(\|x\|,\|y\|)$.

LEMMA 2†: For any compact convex set H in E^N, there is a positive number 1 such that for all $x \in H^*-H_g^*$, there is an $x'' \in H_g^*$, such that

$$x''>x \text{ and } f(x''-x)=1.$$

PROOF: For any $x \in H^*-H_g^*$, let $F(x)$ be the union of the single point x with the set of all $x' \in H^*$ satisfying (2) $F(x)$ is compact, and therefore the function $\|y-x\|$, considered as a function of y, attains its maximum in $F(x)$, say at the point x''. Suppose $x'' \notin H_g^*$, then by lemma 1 there is a point $y \in H^*$ such that $y>x''$ and $f(y-x'')=1$.

From (1) since $y-x=(y-x'')+(x''-x)$, $f(y-x'')=1$ and $f(x''-x)=1$ it then follows that $f(y-x)=1$ and from (3)

$$\text{that } ||y-x||=||(y-x'')+(x''-x)||>\max(||y-x''|| \quad ||x''-x||) \geqslant x''-x$$

hence x'' does not have the maximum property by which it was defined, which is a contradiction. This completes the proof.

COROLLARY 3††: If H is a compact convex subset of E^N, then for all $Z \in H^*-H_g^*$, there is a $\hat{Z} \in H_g^*$ such that $\hat{Z}>Z$ and for each $i \in N$, $\hat{Z}^i-Z^i>||\hat{Z}-Z||$.

THEOREM: Assume that H is a convex compact subset of E^N. Then

$$C(H^*)=C(H_i^*)=C(H_g^*)=C(H_{ig}^*).$$

PROOF: The difficult part is to prove that imposing group rationality, either on H^* or H_i^*, does not change the core. For example, take H^*, we must prove $C(H^*)=C(H_g^*)$. $C(H^*) \subset \smile(H_g^*)$ is easily established. The crux of the proof is $C(H_g^*) \subset C(H^*)$. Let $x \in C(H_g^*)$. We will suppose $x \notin C(H^*)$, i.e., x is dominated via some S by a payoff vector y of H^*, and will then construct an element $\hat{Z}$ of H_g^* which also dominate x; this will be a contradiction.

More precisely, we can assume without loss of generality that $x=0$. Let $6=\min_{i \in S} y^i$, where $y \succ_S x$. Let's take an element Z very close to x on the line segment $\bar{x}\bar{y}$ which joins x and y, that is, let $Z \in \bar{y}\bar{x}$ be such that $|Z|<\delta/2$. Then,

$$Z \in H^*-H_g^*.$$

†This lemma corresponds to lemma 4 in [2].

††This corollary corresponds to corollary 5 in [2], and lemma in [1].

By corollary 3 there is an $\hat{Z} \in H_g^*$ such that $\hat{Z} > Z$ and

$$\hat{Z}^i - Z^i > ||\hat{Z} - Z|| \text{ for all } i \in N.$$

Hence if $||\hat{Z} - Z|| > \delta/2$, then $\hat{Z}^i - Z > \delta/2$. Then,

$$\hat{Z}^i = \hat{Z}^i - Z^i + Z^i \geq \hat{Z}^i - Z^i - ||Z|| > \delta/2 - \delta/2 = 0,$$

this contradicts $0 \in H_g^*$. Hence $||\hat{Z} - Z|| < \delta/2$. Hence for all $i \in N$,

$$\hat{Z}^i = \hat{Z}^i - Z^i + Z^i \leq ||\hat{Z} - Z|| + ||Z|| < \delta/2 + \delta/2 = \delta.$$

Hence for all $i \in S$,

$$\hat{Z}^i - y^i \leq \max_i \hat{Z}^i - \min_i y^i < \delta - \delta = 0.$$

Hence $\hat{Z}^s < y^s$. Therefore $\hat{Z}^s \in v(s)$, and $\hat{Z} \prec_s x$. This contradicts $x \in C(H_g^*)$. This completes the proof.

4. In an ordinary game in the sense of Aumann [2], H is the set of vectors such that N can get exactly x. And we are interested in the core of H instead of H^*.

If H is a compact convex subset of E^N such that $H = H^*$, then $C(H) = C(H^*)$. So we suggest the axioms that will make the theory even more satisfying.

DEFINITION 4.1†: An n-person characteristic function is a pair (N, v) where N is a set with n players, and v is a function that carries each subset $S \subset N$ into a subset $v(s) \subset E^N$ so that

(1) $v(s)$ is convex;

(2) $v(s)$ is closed;

(3) $v(\phi) = E^N$; and

(4) v is a comprehensive, i.e., if $x \in v(S)$ and $x^S \geq y^S$ then $y \in v(S)$.

DEFINITION 4.2: An n-person game is triple (N, v, H), where (N, v) is a characteristic function, and

(5) H is a compact convex subset of E^N. We shall say that (N, v, H) is an ordinary game if

(6) $x \in v(N)$ if and only if there is a $y \in H$ such that $x \leq y$; and

(7) $v(N) - H$ is a comprehensive, i.e., if $x \in v(N) - H$ and $x \geq y$, then $y \in v(N) - H$.

Now as defined in section 3, $H^* = H_+ \cap H_-$, where, $H_+ = \cup_{x \in H} H + (x)$.

$H_+(x) = \{y;\ y \in E^N$, for $a\ x \in H$ such that $y \leq x\}$, so that we have

$$H_+ = \{x;\ \text{there is a } y \in H \text{ such that } x \leq y\}.$$

In the same way, $H_- = \{x;$ there is a $y \in H$ such that $x \geq y\}$.

Therefore, $H_+ = v(N)$.

Condition (7) is equivalent to $H = H^*$.

That is,

LEMMA 4.3: $v(N) - H$ is comprehensive if and only if $H = H^*$.

PROOF: Let $v(N) - H$ be a comprehensive. We must prove $H = H^*$.

Clearly $H \subset H^*$. The crux of the proof is $H^* \subset H$. Let $x \in H^*$, suppose $x \notin H$.

$$H^* = H_+ \cap H_- \subset H_+ = v(N).$$

---

†In this definition v is not assumed to be super-additive, i.e., the condition $v(S \cup T) \supset v(S) \cap v(T)$ for every pair of disjoint coalition S and T is dropped.

Hence $x \in v(N)$ and $x \notin H$. Hence $x \in v(N) - H$.

$$H^* = H_+ \cap H_- \subset H_-, \text{ hence } x \in H_-.$$

As $x \in H_-$ there is a $y \in H$ such that $x \geqslant y$. Since $v(N) - H$ is a comprehensive, we obtain $y \in v(N) - H$. That is, $y \in v(N)$ and $y \notin H$. But $y \in H$, this is a contradiction. Now, let $H = H^*$. Let $x \in v(N) - H$ and $y \leqslant x$. Then $x \in v(N)$ and $x \notin H$ and $y \leqslant x$. By (4) $x \in v(N)$ and $y \leqslant x$ imply $y \in v(N)$. Next we show that $x \notin H$ and $y \leqslant x$ imply $y \notin H$. Let $x \notin H$ and $y \leqslant x$, suppose $y \in H$. Then $y \in H$, $H = H^*$, and $H^* \subset H_-$ imply $y \in H_-$. Since $H_- = \{y;$ there is a $Z \in H$ such that $y \geqslant Z\}$, $y \in H_-$ and $y \leqslant x$ imply $x \in H_-$. Hence $x \in v(N)$ and $v(N) = H_+$ imply $x \in H_+$. Hence $x \in H^*$. Since $H = H^*$, $x \in H$. This is a contradiction. After all, $y \in v(N)$ and $y \notin H$, that is,

$$y \in v(N) - H.$$

This completes the proof. Then, we obtain a theorem of the core.

THEOREM 4.3. (N, v, H) is an ordinary game, then

$$C(H) = C(Hi) = C(Hg) = C(Hig).$$

That is, all the "interesting" cores coincide.

PROPOSITION 4.4. V is H_i-stable if and only if V is H_{ig}-stable.

PROOF: Peleg in [4] proved the proposition under the assumption of (1), (2), (3), (4) and (5′) H is a compact convex subset of $v(N)$ instead of (5). In the simultaneous way, we can prove the proposition. So, we omit the proof.

DEFINITION 4.5: Let $G = (N, v, H)$ be an n-person game. A solution of G is an H_{ig}-stable set.

5. From the above discussion we can understand that the core theorem is true in the case H is not a polyhedron. And in counter-example by Aumann [2], $C(H)$ does not coincide with $C(H_g)$ because $H \subsetneq H^*$ in view of our axioms. In this example we obtain from H that H^* is the convex hull of the sets C' and D, where

$$C' = \{x: x^1 \geqslant 0, x^2 \geqslant 0, x^3 = 0, (x^1)^2 + (x^2 + 1)^2 \leqslant 4\}$$

$$D = \{x: x^1 \geqslant 0, x^2 \geqslant 0, x^3 = 1, (x^1)^2 + (x^2 + 1)^2 \leqslant 4\}.$$

Therefore, clearly $C(H^*) = C(H_g^*)$.

In an ordinary game in the sense of Aumann [2], H is the set of vectors such that N can get exactly x. So H should satisfy $H = H^*$.

REFERENCES

[1] Aumann, R. J. "A Survey of Cooperative Games without Side Payments," *Essays in Mathematical Economics, In Honour of Oskar Morgenstern* (Edited by Martin Shubik) (Princeton Univ. Press, Princeton, New Jersey, 1967), pp. 3–27.

[2] Aumann, R. J. "The Core of a Cooperative Game without Side Payments," Trans. Am. Math. Soc. *98*, 539–552 (1961).

[3] Aumann, R. J. and B. Peleg, "Von Neumann-Morgenstern Solutions to Cooperative Games without Side Payments," Bull. Am. Math. Soc. *66*, 173–179 (1960).

[4] Peleg, B. "Solutions to Cooperative Games without Side Payments," Trans. Am. Math. Soc. *106*, 280–292 (1963).

[5] Stearns, R. E. "Three-Person Cooperative Games without Side Payments," Annals of Mathematic Studies *52*, 377–406 (1964).

[6] von Neumann, J. and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton Univ. Press, Princeton, N.J., 1953, 3rd. ed.

* * *

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Washington, D.C. 20360. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

MARCH 1970
VOL. 17, NO. 1
NAVSO P-1278

CONTENTS

ARTICLES

	Page
Machine Sequencing: Disjunctive Graphs and Degree-Constrained Subgraphs by E. Balas	1
Review of Sequencing Research by J. E. Day and M. P. Hottenstein	11
Computer-Aided Selection of Prior Distributions for Generating Monte Carlo Confidence Bounds on System Reliability by N. R. Mann	41
A Sequential Bayes Procedure for Reliability Demonstration by R. E. Schafer and N. D. Singpurwalla	55
Confidence Intervals for Ranked Means by E. J. Dudewicz	69
A Two-Echelon Multi-Station Inventory Model for Navy Applications by S. Zacks	79
A Game Theory Approach to Logistics Allocation by S. Moglewer and C. Payne	87
Multiple Dispatches in a Poisson Process by C. E. Bell	99
A Comparative Study of Demand Forecasting Techniques for Military Helicopter Spare Parts by R. E. Markland	103
Stochastic Duels with Nonrepairable Weapons by N. Bhashyam	121
Suboptimal Ordering Policies Under the Full Cost Criterion by S. C. Jaquette	131
Operational Definitions of Inventory Record Accuracy by D. A. Schrady	133
A Note on the Core of a Cooperative Game without Side Payment by Y. Yasuda	143

OFFICE OF NAVAL RESEARCH
Washington, D.C. 20360